

SALIENCY TUBES: VISUAL EXPLANATIONS FOR SPATIO-TEMPORAL CONVOLUTIONS

Alexandros Stergiou^{1†}, Georgios Kapidis^{1,4}, Grigorios Kalliatakis², Christos Chrysoulas³, Remco Veltkamp¹, Ronald Poppe¹

¹ Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands

² School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom

³ School of Computer Science and Informatics, London South Bank University, London, United Kingdom

⁴ Noldus Information Technology, Wageningen, The Netherlands

{a.g.stergiou, g.kapidis, r.c.veltkamp, r.w.poppe}@uu.nl, gkallia@essex.ac.uk, chrysouc@lsbu.ac.uk

ABSTRACT

Deep learning approaches have been established as the main methodology for video classification and recognition. Recently, 3-dimensional convolutions have been used to achieve state-of-the-art performance in many challenging video datasets. Because of the high level of complexity of these methods, as the convolution operations are also extended to an additional dimension in order to extract features from it as well, providing a visualization for the signals that the network interpret as informative, is a challenging task. An effective notion of understanding the network's inner-workings would be to isolate the spatio-temporal regions on the video that the network finds most informative. We propose a method called *Saliency Tubes* which demonstrate the foremost points and regions in both frame level and over time that are found to be the main focus points of the network. We demonstrate our findings on widely used datasets for third-person and egocentric action classification and enhance the set of methods and visualizations that improve 3D Convolutional Neural Networks (CNNs) intelligibility. Our code ¹ and a demo video ² are also available.

Index Terms— Visual explanations, explainable convolutions, spatio-temporal feature representation.

1. INTRODUCTION

Deep Convolutional Neural Networks (CNNs) have enabled unparalleled breakthroughs in a variety of visual tasks, such as image classification [1, 2], object detection [3], image captioning [4, 5], and video classification [6, 7, 8]. While these deep neural networks show superior performance, they are often criticized as black boxes that lack interpretability, because of their end-to-end learning approach. This hinders the understanding of which features are extracted and what improvements can be made in the architectural level.

[†]Corresponding author

¹<https://goo.gl/xX4nnv>

²<https://youtu.be/JANUqoMc3es>

Hence, there has been a significant interest over the last few years in developing various methods of interpreting CNN models [9, 10, 11, 12]. One such category of methods probes the neural network models by trying to change the input and analyzing the model's response to it. Another approach is to explain the decision of a model by training another deep model which reveals the visual explanations.

While there has been promising progress in the context of these 'visual explanations' for 2D CNNs, visualizing learned features of 3D convolutions, where the networks have access to not only the appearance information present in single, static images, but also their complex temporal evolution, has not received the same attention. To extend 'visual explanations' to spatio-temporal data such as videos, we propose Saliency Tubes, a generalized attention mechanism for explaining CNN decisions, which is inspired by the class activation mapping (CAM) proposed in [13].

Saliency Tubes is a general and extensible module that can be easily plugged into any existing spatio-temporal CNN architecture to enable human-interpretable visual explanations across multiple tasks including action classification and egocentric action recognition.

Our key contributions are summarized as follows:

- We propose Saliency Tubes, a spatio-temporal-specific, class-discriminative technique that generates visual explanations from any 3D ConvNet without requiring architectural changes or re-training.
- We apply Saliency Tubes to existing top-performing video recognition spatio-temporal models. For action classification, our visualizations highlight the important regions in the video for predicting the action class, and shed light on why the predictions succeed or fail. For egocentric action recognition, our visualizations point out the target objects of the overall motion, and how their interaction with the position of the hands indicates patterns of everyday actions.
- Through visual examples, we show that Saliency Tubes improve upon the region-specific nature of Map meth-

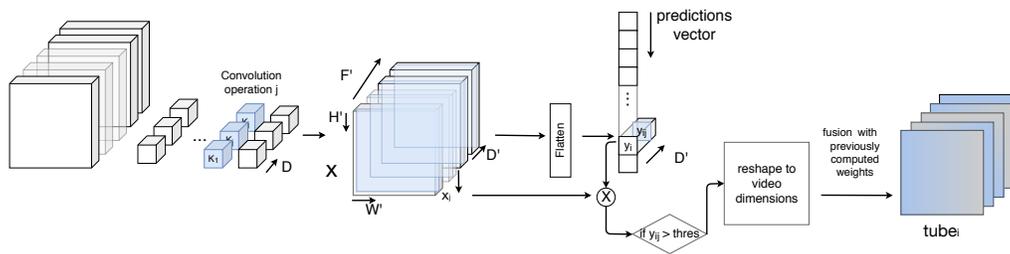


Fig. 1. Saliency Tubes. (a) Informative regions are found based on the activation maps from output x , while the useful features are defined based on their corresponding values in feature vector y_i . (b) Individual features can also be visualised spatio-temporally by fusing all the activation maps in the last step of the focus tubes and only include single re-scaled activation maps. The illustration is presented for the simplified case of a single convolution layer for convenience and easier interpretability.

ods by showing a generalized spatio-temporal focus of the network.

Related work on visual interpretability for neural network representations is summarized in Section 2. In Section 3 the details of the proposed approach are presented. In Section 4, we report the visualization results in third person and egocentric action classification and we discuss their descriptiveness. The paper’s main conclusions are drawn in Section 5.

2. RELATED WORK

Bau *et al.* argue [14], that two of the key elements of CNNs should be their discrimination capabilities and their interpretability. Although the discrimination capabilities of CNNs have been well established, the same could not be said about their interpretability as ways of visualizing each of their binding parts have been proven challenging. The direct visualization of convolutional kernels has been a well explored field with many works based on the inversion of feature maps to images [15] as well as gradient-based visualizations [16, 17]. Following the gradient-based approaches, one of the first attempts to present the network’s receptive field was proposed by Zhou *et al.* [18] in which the output neural activation feature maps were represented with image-resolution. Others have also focused on parts of networks and how specific inputs can be used to identify the units that have larger activations [19]. Konam [20] proposed a method for discovering regions that excite particular neurons. This notion was later used as the main point for creating explanatory graphs, which correspond to features that are tracked though the CNN’s extracted feature hierarchy [21, 22] and are based on separating the different feature parts of convolutional kernels and representing individually the parts of different extracted kernel patterns.

Only few works have addressed the video domain, aimed at reproducing the visual explanations achieved in image-based models. Karpathy *et al.* [23] visualized the Long Short Term Memory (LSTM) cells. Bargal *et al.* [24] have studied

class activations for action recognition in systems composed of 2D CNN classifiers combined with LSTMs for monitoring the temporal variations of the CNN outputs. Their approach was based on Excitation Backpropagation [25]. Their main focus was based on the decision made by Recurrent Neural Networks (RNNs) in action recognition and video captioning, with the convolutional blocks only being used as per-frame feature extractors. In 3D action recognition, Chattopadhyay *et al.* [9] have proposed a generalized version of the class activation maps generalized for object recognition.

To address the lack of visual explanation methods for 3D convolutions, we propose Saliency Tubes, which are constructed for finding both regions and frames the network focuses on. This method can be generalized to different action-based approaches in videos as we demonstrate for both third person and egocentric tasks.

3. SALIENCY TUBES

Figure 1 outlines our approach. We let x denote the activation maps of the network’s final convolutional layer with output maps of size $F' \times W' \times H' \times D'$, where F' represents the number of frames that are used, W' is the width of the activation maps, H' is the height and D' is the number of channels (also could be referred to as frame-wide depth) that equal the total number of convolutional operations performed in that layer. Let also y_i be the tensor in the final fully-connected layer responsible for class predictions, of a specific class i (with $i \in \{0, N\}$ and N being the total number of classes). We consider every element of the predictions vector, denoted as $y_{i,j}$ which corresponds to a specific depth dimension of the network’s final convolutional layer (x) and designates how informative that specific activation map is towards a correct prediction for an example of class i . In order to do so, we propagate back to these activation maps ($a_{f,w,h,j}$) and multiply all their elements by the equivalent predictions weight vector $y_{i,j}$. The class weighted operation can be formulated as $z_{i,j}$ in which:

$$z_{i,j} = \sum_f^{F'} \sum_w^{W'} \sum_h^{H'} y_{i,j} \times a_{f,w,h,j} \quad \forall y_{i,j} \geq \tau \quad (1)$$

Because of the large number of features that are extracted by the network (dimension D' could take a value in the range of thousands in modern architectures), we specify a threshold τ based on which, only the activations that significantly contribute to the predictions are selected. We define all values below this threshold as elements of set E .

Following the matrix multiplication to find the feature's intensity, the activations are then reshaped to correspond to the original video dimensions of $F \times W \times H$. During the reshape process we use spline interpolation for increasing the spatio-temporal dimensions of the thresholded activations. To create the final saliency tubes, the operation described in Equation 1 is performed for j features with the final output being:

$$tube_i = \sum_j^{D'} z_{i,j}, \quad \forall j \in \{0, D'\} \notin E. \quad (2)$$

4. VISUALIZATION OF SALIENCY TUBES

In Section 4.1 we visualize the outputs of 3D CNNs using Saliency Tubes in two different forms and in Section 4.2 we compare them against the outputs of 2D CNNs.

4.1. Localization of Saliency Tubes

In Figure 2, we demonstrate two cases of video activity classification with overlaid Heat and Focus Tubes, which we utilize as a means to visualize the Saliency Tubes. To produce the activation maps we use a 3D Multi-Fiber Network (MFNet) [26] pretrained on Kinetics [27] and subsequently finetuned on UCF-101 [28] and EPIC-Kitchens (verbs) [29], respectively. Our aim is to examine the regions in space and time that the networks focus on for a particular action class and feature. In our examples, the network input is 16 frames, which we also use as a visualization basis to overlay the output.

In row 1 of Figure 2, we show the example of a person performing a martial arts exhibition (TaiChi class), from the test set of UCF-101 [28]. The Saliency Tubes show that the network does not fixate on the person but follows parts that correlate with the movement as it progresses. We observe high activations during the backstep and left-hand motions, but not the front-step in between. This shows that the network finds some specific action segments more informative than others, in terms of selected classes instead of the whole range of motions that exist in the video.

In Fig. 2 row 2, we visualize a segment from the EPIC-Kitchens [29] dataset with the action label 'open door'. Here,

Original video Heat Tube Focus Tube

Fig. 2. Visualizing Saliency Tubes. Row 1 presents examples from 3rd person perspective videos, such as those found in UCF-101 [28]. For the second row we focus on egocentric tasks from the EPIC-Kitchens dataset [29]. For both examples, we use a 3D Multi-Fiber Network [26], pre-trained on the Kinetics dataset [27] and finetuned on each of the two datasets. Viewed better on Adobe Reader where the subfigures play as videos.

our classifier is trained only on verb classes, therefore we expect it to consider motions as more significant than appearance features when predicting a class label. Initially, the moving hand is shown to produce relatively high activations; significantly higher compared to the 'door' area which is the main object of the segment. After a period of movement towards the door that is not considered meaningful, high activations are correlated with the door's movement. This leads to the realization that the network notices this movement and takes these features into account for the class prediction. It is important to note that the focus of the activations does not depend solely on the moving object, but is largely dependant on the area of the motion. Finally, as the door moves out of the scene, the activations remain high in the area in which it used to be. This analysis of a 3D network's output is only possible due to Saliency Tube's ability to visualize its activations as a whole and not per frame.

4.2. Saliency comparison of 2D and 3D Convolutions

We further compare our results to the ones obtained by directly using 2D convolutions. More specifically, we use a Temporal Segment Network (TSN) [30] pre-trained on ImageNet [2] and finetuned on EPIC-Kitchens (verbs) [29] to demonstrate the class activations in videos of 2D convolutions, while we use the MFNet [26] from our previous example for spatio-temporal activations. In Figure 3, a bounding box annotation from [29] is regarded as the possible area of interest in the scene which we overlay on the corresponding heat-maps [11] and heat-tubes of the final convolutional layer from each network respectively. The heat-maps created were based on slight modifications in our method in order to correspond with the decreased tensor dimensionality.

The 2D convolutions from TSN show time-invariant activations, meaning that the model will make class predictions based on appearance features in every frame. Therefore, the movement occurring in the action is not taken into account, making the predictions to depend heavily on both model complexity (as for overfitting) and strong inter-class similarities. This also empowers the notion of using supplementary crafted temporal features (such as optical flow) for including motion features as input for the network. In contrast, Saliency Tubes exhibit that temporal movement is highly influential to 3D convolutions when determining class features. Our visualizations confirm that alongside finding regions in each frame where class features are present, 3D CNNs also reveal the frames in which these features are present in greater concentration.

Fig. 3. Comparison between 2D and 3D saliency. The main action of the video is *stirring* and it primarily takes place in the middle of the clips. 2D convolutions (left) focus significantly on object appearance without taking into consideration the movements that are performed in the video. This can be seen as every frame in the case of 2D convolutions includes some feature activation. In contrast, 3D convolutions (right) only extract image regions in specific frames where motions are present.

5. CONCLUSIONS

In this work, we propose Saliency Tubes as a way to visualize the activation maps of 3D CNNs with relation to a class of interest. Previous work on 2D CNNs establishes visualization methods as a way to increase interpretability of convolutional neural networks and as a supplementary feedback mechanism in terms of dataset overfitting. We build upon this idea for 3D convolutions, using a simple yet effective concept that represents regions in space and time in which the network locates the most discriminative class features.

Additionally, using our visualization scheme, we further validate the notion that 3D convolutions are more effective in learning motion-based features from temporal structures, and they do not only include a larger number of tensor parameters that allow them to achieve better results. We support this by demonstrating how a 2D CNN will focus only on the appearance features per frame for its prediction, whereas a 3D CNN produces a more elaborate spatio-temporal analysis.

6. REFERENCES

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 770–778.
- [2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “Imagenet classification with deep convolutional neural networks,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097–1105.
- [3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 580–587.
- [4] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, “Microsoft COCO captions: Data collection and evaluation server,” *arXiv preprint arXiv:1504.00325*, 2015.
- [5] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, “Show and tell: A neural image caption generator,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 3156–3164.
- [6] Karen Simonyan and Andrew Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.
- [7] Georgia Gkioxari and Jitendra Malik, “Finding action tubes,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CPVR)*. IEEE, 2015, pp. 759–768.
- [8] Alexandros Stergiou and Ronald Poppe, “Understanding human-human interactions: A survey,” *arXiv preprint arXiv:1808.00022*, 2018.
- [9] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.
- [10] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller, “Methods for interpreting and understanding deep neural networks,” *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

- [11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 618–626.
- [12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, “Why should I trust you?: Explaining the predictions of any classifier,” in *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*. ACM, 2016, pp. 1135–1144.
- [13] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, “Learning deep features for discriminative localization,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2921–2929.
- [14] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba, “Network dissection: Quantifying interpretability of deep visual representations,” *arXiv preprint arXiv:1704.05796*, 2017.
- [15] Alexey Dosovitskiy and Thomas Brox, “Inverting visual representations with convolutional networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 4829–4837.
- [16] Matthew D Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” in *Proceedings of the European conference on computer vision, (ECCV)*. Springer, 2014, pp. 818–833.
- [17] Aravindh Mahendran and Andrea Vedaldi, “Understanding deep image representations by inverting them,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 5188–5196.
- [18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, “Object detectors emerge in deep scene CNNs,” *arXiv preprint arXiv:1412.6856*, 2014.
- [19] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson, “Understanding neural networks through deep visualization,” *arXiv preprint arXiv:1506.06579*, 2015.
- [20] Sandeep Konam, *Vision-based navigation and deep-learning explanation for autonomy*, Ph.D. thesis, Masters thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, 2017.
- [21] Quanshi Zhang, Ruiming Cao, Feng Shi, Ying Nian Wu, and Song-Chun Zhu, “Interpreting CNN knowledge via an explanatory graph,” in *AAAI Conference on Artificial Intelligence*, 2018.
- [22] Quanshi Zhang, Ruiming Cao, Ying Nian Wu, and Song-Chun Zhu, “Growing interpretable part graphs on convnets via multi-shot learning,” in *AAAI*, 2017, pp. 2898–2906.
- [23] Andrej Karpathy, Justin Johnson, and Li Fei-Fei, “Visualizing and understanding recurrent networks,” *arXiv preprint arXiv:1506.02078*, 2015.
- [24] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff, “Excitation backprop for RNNs,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 1440–1449.
- [25] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff, “Top-down neural attention by excitation backprop,” *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, 2018.
- [26] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng, “Multi-fiber networks for video recognition,” in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 352–367.
- [27] Joao Carreira and Andrew Zisserman, “Quo vadis, action recognition? A new model and the Kinetics dataset,” in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4724–4733.
- [28] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint arXiv:1212.0402*, 2012.
- [29] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al., “Scaling egocentric vision: The EPIC-KITCHENS dataset,” in *Proceedings of the European Conference of Computer Vision (ECCV)*. Springer, 2018.
- [30] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2016, pp. 20–36.