16
Understanding Patch-Based Learning of Video Data by Explaining Predictions

Christopher J. Anders[1], Grégoire Montavon[1], Wojciech Samek[2], and Klaus-Robert Müller[1,3,4]

[1] Technische Universität Berlin, 10587 Berlin, Germany
{gregoire.montavon,klaus-robert.mueller}@tu-berlin.de
[2] Fraunhofer Heinrich Hertz Institute, 10587 Berlin, Germany
wojciech.samek@hhi.fraunhofer.de
[3] Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea
[4] Max Planck Institute for Informatics, Saarbrücken 66123, Germany

Abstract. Deep neural networks have shown to learn highly predictive models of video data. Due to the large number of images in individual videos, a common strategy for training is to repeatedly extract short clips with random offsets from the video. We apply the deep Taylor / Layer-wise Relevance Propagation (LRP) technique to understand classification decisions of a deep network trained with this strategy, and identify a tendency of the classifier to look mainly at the frames close to the temporal boundaries of its input clip. This "border effect" reveals the model's relation to the step size used to extract consecutive video frames for its input, which we can then tune in order to improve the classifier's accuracy without retraining the model. To our knowledge, this is the first work to apply the deep Taylor / LRP technique on any neural network operating on video data.

Keywords: Deep Neural Networks · Video Classification · Human Action Recognition · Explaining Predictions

16.1 Introduction

Deep neural networks have set new standards of performance in many machine learning areas such as image classification [12, 32], speech recognition [6, 19], video analysis [9, 10], or in the sciences [2, 31, 25]. For applications where the input signal is very large in time or space, it has been a common practice to train the model on small patches or clips of that signal [9, 7, 3]. This strategy reduces the number of input variables to be processed by the network and thus, allows to extract the problem's nonlinearities more quickly by performing more training iterations.

An underlying assumption of patch- or clip-based training is the locality of the label information. This assumption is often violated in practice: For example, discriminative information may only be contained in long-term interactions

[6, 36, 19] or only reside at specific time steps (e.g. when a particular action occurs). Since such label noise makes the training more difficult [22], recent work investigated ways to cope with this problem, e.g., attention mechanisms [27] or weighted patch aggregation [3].

This paper aims to investigate patch- or clip-based learning from another perspective, namely by analyzing the properties of a model trained with this specific learning procedure. One way to study the properties of a model is to perform introspection into how the model predicts, for example, by explaining its predictions in terms of input variables [21]. Such explanations can now be robustly obtained for a wide range of convolution-type or general deep neural networks [37, 29, 35, 1, 23, 17, 14], and other machine learning models (e.g., [11]).

In this work we analyzed a convolutional neural network [33] trained for human action recognition on the Sports1M dataset [10] using the deep Taylor / Layer-wise Relevance Propagation (LRP) decomposition technique [17, 15]. We first show that this explanation technique reliably captures class-relevant information from videos. We then test how clip-based training affects the prediction strategy of the network and identify two effects induced by this training procedure. The "border effect" describes the observation that the prediction is predominantly focused at the frames close to the temporal boundaries of its given input to compensate for a small amount of frames per input video clip, whereas the "lookahead effect" describes the observation that the model learns to ignore the first few frames of the input video clip and assign more relevance to the later ones. Finally we demonstrate that the insights obtained by explaining predictions can be directly (i.e. without retraining) used to increase the prediction accuracy of the classifier.

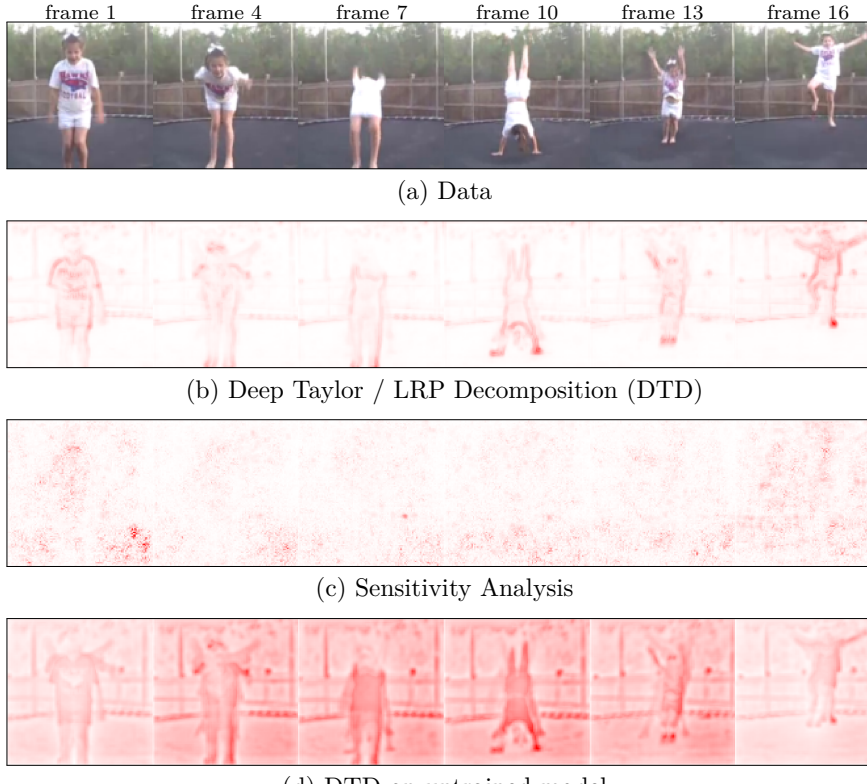
While a different approach for human action recognition has been analyzed before [30] using the LRP framework [1], to our knowledge this work is the first to analyze any neural network for video classification using the deep Taylor / LRP decomposition technique [17]. In a recent work, voxel explanations of 3D-CNNs [34] have been produced using different explanation frameworks [39, 26]. Further research has been done on the interpretation [4], description [38] and segmentation of videos [20]. Outside the field of machine learning, some work has been done on saliency detection in videos [8, 16].

16.2 Explaining the Classifier's Predictions

In this paper, we use the deep Taylor / LRP decomposition technique [17] to produce explanations. We give a brief textual description of the method, along with connections to previous work. The method performs a sum-decomposition of the function value $f(x)$ in terms of input variables [21]

$$f(x) = \sum_{p,t} R_{p,t} \tag{16.1}$$

where $R_{p,t}$ is the relevance of pixel p in frame t. These scores are obtained by progressively redistributing the output $f(x)$ backwards in the network,



(a) Data

(b) Deep Taylor / LRP Decomposition (DTD)

(c) Sensitivity Analysis

(d) DTD on untrained model

Fig. 16.1. Example of a video along with the DTD explanation of this video belonging to the class 'Tumbling'. High relevance scores are shown in red.

differently from back-propagation, until the input variables are reached. This redistribution procedure satisfies a conservation principle [13, 1], where each neuron passes to the lower-layer as much as it has received from the higher layer. Let i, j be neurons of adjacent layers. Let a_i be the activation of neuron i and w_{ij} be the weight that connects it to neuron j. In linear layers, the redistribution is in proportion to the positive contribution of the input activations $R_{i \leftarrow j} \propto a_i w_{ij}^+$ of each neuron [1, 17]. In pooling layers, the redistribution is in proportion to the activations a_i inside the pool [17]. For the first convolutional layer we redistribute in proportion to the signed contributions plus some additive term $R_{i \leftarrow j} \propto a_i w_{ij} - l_i w_{ij}^+ - h_i w_{ij}^-$ where l_i and h_i are the minimal and maximal pixel values respectively [17].

Another popular explanation technique is sensitivity analysis [5, 28], which computes importance scores as e.g.

$$S_{p,t} = \left(\frac{df}{dx_{p,t}} \right)^2. \tag{16.2}$$

We note that this analysis can be interpreted as performing a sum-decomposition of the squared gradient norm ($\|\nabla f\|^2 = \sum_{p,t} S_{p,t}$), and is thus closer to an explanation of the function's variation. We refer the reader for a comparison of different explanation methods to [24, 18].

16.3 Experiments

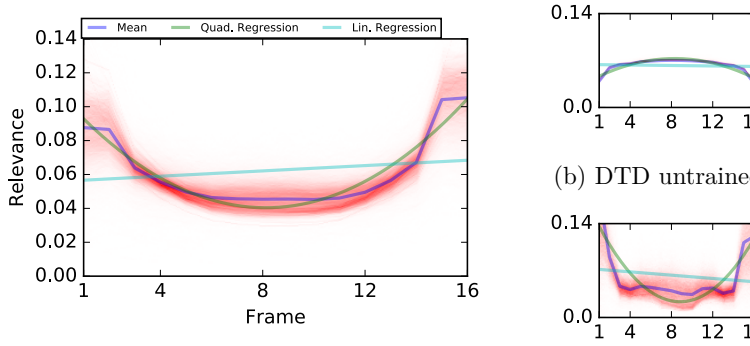
We use the 3-dimensional convolutional neural network architecture C3D as described by [33], with $1+1+2+2+2$ convolutional layers, each group followed by a max-pooling layer and finally 2 consecutive dense layers, where each linear layer is followed by a ReLU activation. Kernel sizes for all convolutions are $3 \times 3 \times 3$, pooling kernels are $1 \times 3 \times 3$ where the dimensions correspond to time by height by width. The network is trained on the Sports-1M data set, which consists of roughly 1 million sports videos from YouTube with 487 classes [10]. Videos are pre-processed by spatially resizing to 128×171 pixels and then center-cropping to 121×121 pixels. We take video clips at particular offsets, composed of 16 frames each. The pre-trained model we use, as supplied by [33], used to be state-of-the-art in human action recognition. It achieves a top-1 accuracy (most confidently predicted class is the label) per clip of 46.1%, a top-1 accuracy over 10 random clips of a single video of 61.1%, as well as a top-5 accuracy (label is in the 5 most confidently predicted classes) for the same setting of 85.2%. Thus, the model successfully performs the classification task and can be analyzed.

We explain predictions for 1000 videos from the test set of Sports-1M using deep Taylor / LRP decomposition [17]. Additional explanations are given for the same 3-dimensional convolutional neural network architecture untrained as well as using gradient-based sensitivity analysis [5, 28] for comparison.

16.3.1 Heatmap Analysis

To get a first impression of the prediction, we take a look at the individual explanation of one specific video clip. In Fig. 16.1, we show an exemplary video and the deep Taylor / LRP decomposition (DTD) for the predicted class label "Tumbling". The hands are identified as relevant, especially when the latter are touching the trampoline, which is characteristic of that class. Other parts of the image such as the trees in the background are not highlighted and therefore found to be non-relevant. The DTD analysis is also less noisy and more focused on the class-relevant features than sensitivity analysis (Fig. 16.1c).

An interesting observation that can be made is that the training procedure tends to make the relevance converge from the center frames of the video clip to its frames closest to the beginning and end respectively as evidenced by the difference between DTD and the same analysis performed on an untrained network (Fig. 16.1d). This so-called border effect will be studied quantitatively in Section 16.3.2. The initial focus on the center of the sequence is due to these frames being more densely connected to the output.



(a) Deep Taylor / LRP Decomposition

(b) DTD untrained

(c) Sensitivity Analysis

Fig. 16.2. Relevance share $(P_t)_t$. Red color shows these vectors for a large number of videos. Lines show the mean relevance share and polynomial fits.

Additional examples of different videos are shown in Figs. 16.6 and 16.7. In particular, the aforementioned observation of higher relevance towards the videos' borders is more clearly visible in Figs. 16.6a and 16.6b. Furthermore, we can also observe that the final frames receive more relevance than any other ones in Figs. 16.6a, 16.7a and 16.7b. This lookahead effect will also be studied quantitatively in Section 16.3.2.

16.3.2 Quantifying Border and Lookahead Effects

To refine the intuition developed in Section 16.3.1 about the presence of a border and lookahead effect, we produce DTD explanations for a large number of videos and analyze their average properties. Because the border effect occurs in the temporal domain, we only focus on the temporal axis of explanations $R_{p,t}$ by defining a frame-wise explanation $R_t = \sum_p R_{p,t}$. From these relevance scores, we can define a vector $(P_t)_t$ where $P_t = R_t / \sum_t R_t$ is the share of relevance at time t. Since our input video clips each contain 16 frames, this vector has size 16, which we can visualize in a plot. Results are shown in Fig. 16.2. The red pattern represents the distribution of these 16-dimensional vectors, for which we can compute an average over the dataset (blue line). Results are also compared to sensitivity analysis, as well as DTD on the untrained model.

Results confirm our previous observations of higher relevance in the bordering frames. Note that DTD and sensitivity analysis (Fig. 16.2c) produce consistent results with respect to the border effect. We can further verify that this effect is not due to an architecture-related artifact, by performing the same DTD analysis on the untrained model (Fig. 16.2b): The border effect is present only for the trained model. For the untrained model, relevance at the border is instead lower compared to other frames. The additional lookahead effect can be observed from

this analysis where the relevance is slightly higher for the last frame as opposed to the first frame.

In order to determine the strength of the border and lookahead effects, we need a quantitative measure for them. We propose to capture these effects by fitting vectors $(P_t)_t$ using simple quadratic regression. More specifically, we consider the quadratic model

$$q(t) = B \cdot t^2 + C \cdot t + D \tag{16.3}$$

and fit the coefficients B, C, D to minimize the least square error $\sum_t \|E[P_t] - q(t)\|^2$, where $E[\cdot]$ is the expectation over the Sports-1M test set. The strength of the border effect is captured by the variable B. Similarly, to capture the lookahead effect, we fit a linear model

$$l(t) = L \cdot t + A \tag{16.4}$$

using similar least squares objective, and identify the lookahead strength by the parameter L. Fitted models $q(t)$ and $l(t)$ are shown as green and cyan lines in Fig. 16.2.

Table 16.1. Parameters for fitted models $q(t)$ and $l(t)$ as in Eqs. 16.3 and 16.4. Relevant coefficients are shown in bold.

	DTD	SA	DTD-u
$\boldsymbol{B}$	**0.0010**	**0.0018**	**−0.0005**
C	−0.0168	−0.0322	0.0082
D	0.1085	0.1661	0.0389
$\boldsymbol{L}$	**0.0007**	**−0.0012**	**−0.0002**
A	0.0558	0.0729	0.0640

These parameters are shown in Table 16.1 for the deep Taylor / LRP decomposition (DTD), sensitivity analysis (SA), and the DTD on the untrained model (DTD-u). Coefficients used for the analysis are shown in bold. We can observe that the border parameter B is positive for both analyses performed on the trained model. The lookahead parameter however has varying signs depending on the choice of analysis. We will see later in Section 16.3.4 that this parameter is influenced by the offset of the input sequence.

16.3.3 Border Effect and Step Size

The border effect can be intuitively understood as an attempt by the network to look beyond the sequence received as input. This suggests that upscaling the input sequence may reduce this effect as more context becomes available. For example, Fig. 16.6a is a static scene with barely any motion and shows, compared to other samples, more relevance at the border frames. To test this,

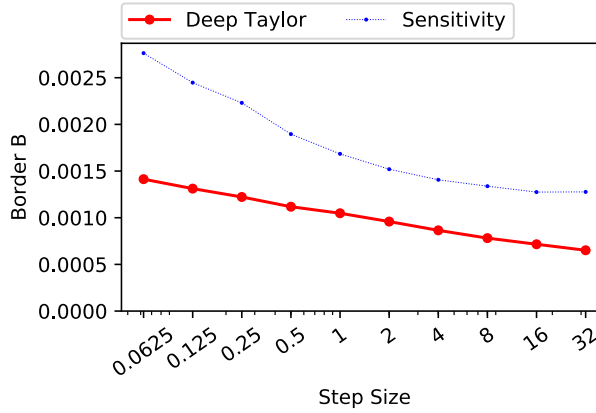


Fig. 16.3. Border parameter B by step size (logarithmic scale)

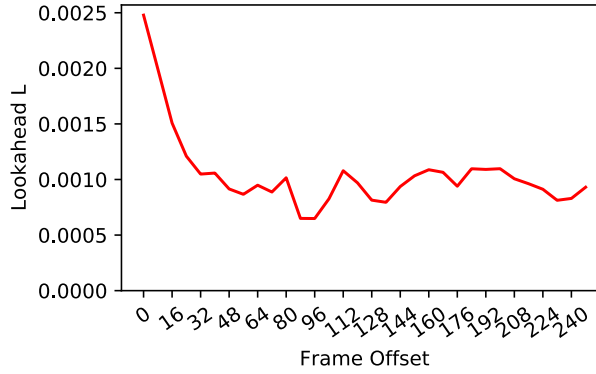


Fig. 16.4. Lookahead parameter L by intra-video frame offset.

we will subsample videos with various step sizes. We start with a step size of $\frac{1}{16}$, which is the same frame repeated 16 times. We then double the step size repeatedly until we reach a value of 32. At each step size, we apply DTD as well as sensitivity analysis. Note that the model is left untouched. The border parameter B for each step size is given in Fig. 16.3. For low step sizes, the border effect is strong. As the step size increases, the border effect is reduced, thus confirming the above intuition.

16.3.4 Lookahead Effect and Offset

The lookahead effect is the tendency of the network to look predominantly at the end of the sequence. We would like to test whether this effect occurs at every position in the video or mainly at the beginning. One of our suspicions is, that many videos start with some opening screen, where the title of the video,

authors etc. are introduced. It would seem natural that the model ignores the first few frames of the video and assigns more relevance to the later frames. An example for such a video clip is shown in Fig. 16.7b. We start by taking the input sequence at the beginning of the video, then, we slide the window by 8-framed offsets until we reach an offset of 256. The results are shown in Fig. 16.4. We observe that for offsets 0, 8 and 16, the lookahead parameter is high compared to other offsets, and becomes low and constant for larger offsets. This behavior for small offsets supports the hypothesis of non-informative content at the beginning of the video.

16.3.5 Step Size and Model Accuracy

As a final experiment, we look at how the step size not only controls the border effect, but also the model's classification accuracy. In particular, we test whether we can improve the classifier accuracy by simply choosing a step size different from the training data, without retraining the model. We use the previously defined step sizes and plot in Fig. 16.5 the resulting border parameter in correspondence to the produced classification accuracy. (The measure of accuracy is the membership of the true label to the top five predictions.) A low step size produces few correct predictions. Performance slowly increases until the highest accuracy is reached at a step size of 2, about 1% above the baseline accuracy of 60%. After that, accuracy drops again until a step size of 16. A key observation here is that the optimal step size is different from the step size 1 used for training the model. Thus, the classification accuracy was improved at no cost. Note that we could have made this observation without any model explanation by simple validation over the frame rate. However, the contribution of the explanation module here is the insight of how the model utilizes frames in the input clip, which led to this experiment in the first place.

16.4 Conclusion

In this work, we have explained the reasoning of a highly predictive video neural network trained on a sports classification task. For this, we have used the recently proposed deep Taylor / LRP framework, which allowed us to robustly identify which frames in the video and which pixels of each frame are relevant for prediction. The method was able to correctly identify video features specific to certain sports. In addition, the analysis has also revealed systematic imbalances in the way relevance is distributed in the temporal domain. These imbalances, that we called "border effect" and "lookahead effect", can be understood as an attempt by the network to look beyond the sequence it receives as input. Based on the result of this analysis, we then explored how transforming the input data reduces/increases these imbalances. In particular, down-sampling the data was shown to reduce the border effect, and also to bring a small increase in classification accuracy (Fig. 16.5), without actually retraining the model. Even though the "lookahead effect" did not immediately lead to a strategy to improve

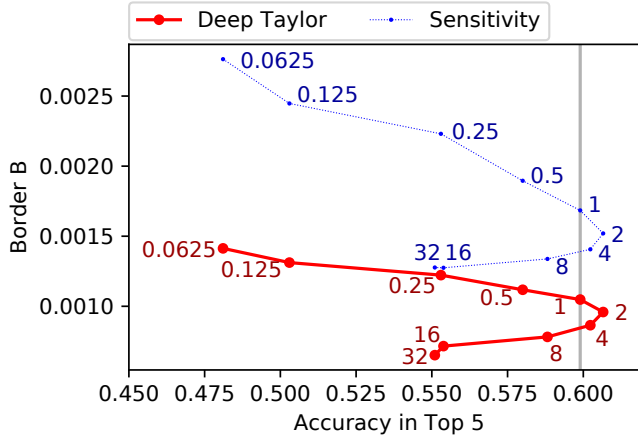


Fig. 16.5. Border parameter B by top-5 accuracy along step size. The grey bar indicates the baseline accuracy.

the model, it implied flaws in the preprocessing of the training data. While these specific findings were only shown in this respective context of C3D and Sports1M, we were able to demonstrate to what extent the findings of such an analysis of a video classifier could be used to gain insight of a model's relation to its input data. We speculate that other models might share similar relations.

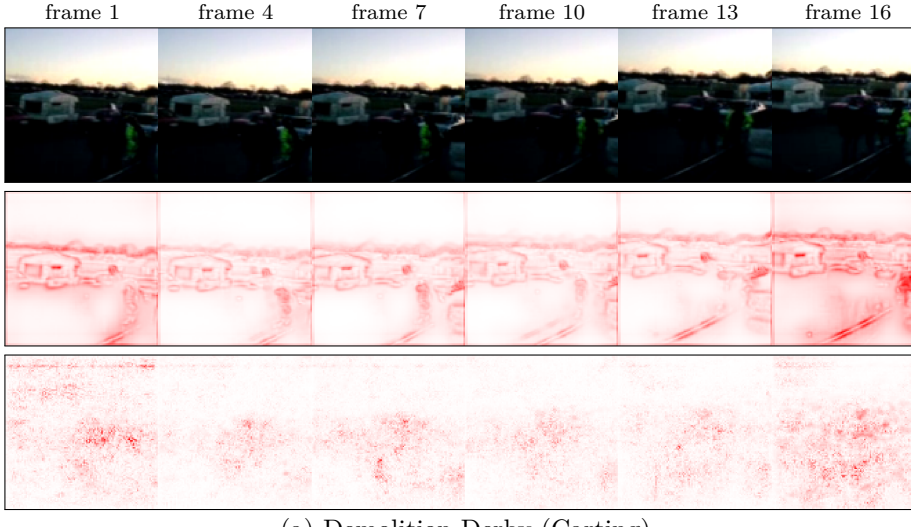
Acknowledgements. This work was supported by the German Ministry for Education and Research as Berlin Big Data Centre (01IS14013A), Berlin Center for Machine Learning (01IS18037I) and TraMeExCo (01IS18056A). Partial funding by DFG is acknowledged (EXC 2046/1, project-ID: 390685689). This work was also supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451, No. 2017-0-01779).

References

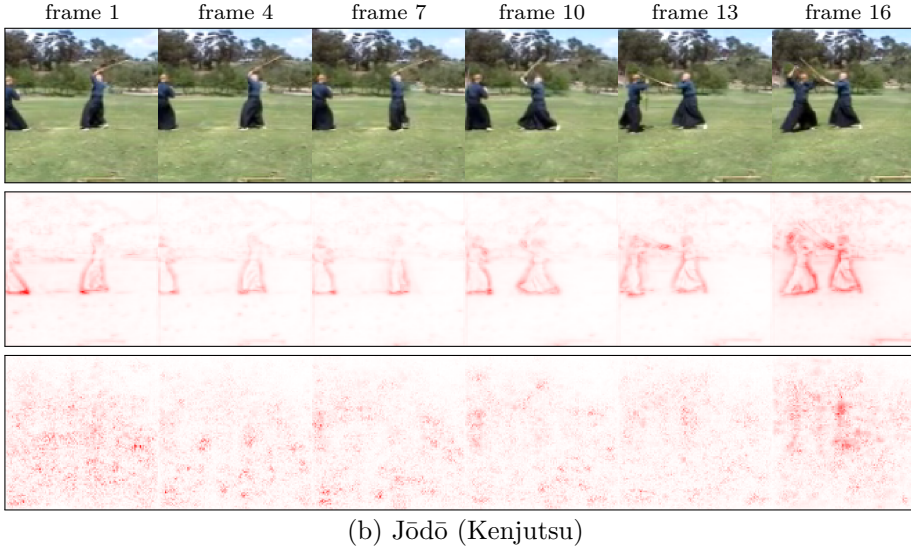
1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015)
2. Baldi, P., Sadowski, P., Whiteson, D.: Searching for exotic particles in high-energy physics with deep learning. Nature communications **5**, 4308 (2014)
3. Bosse, S., Maniry, D., Müller, K.R., Wiegand, T., Samek, W.: Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on Image Processing **27**(1), 206–219 (2018)
4. Donahue, J., A. Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: IEEE CVPR. pp. 2625–2634 (2015)

5. Gevrey, M., Dimopoulos, I., Lek, S.: Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecological Modelling **160**(3), 249–264 (2003)
6. Graves, A., Mohamed, A., Hinton, G.: Speech recognition with deep recurrent neural networks. In: IEEE ICASSP. pp. 6645–6649 (2013)
7. Hou, L., Samaras, D., Kurc, T.M., Gao, Y., Davis, J.E., Saltz, J.H.: Patch-based convolutional neural network for whole slide tissue image classification. In: IEEE CVPR. pp. 2424–2433 (2016)
8. Hu, K.T., Leou, J.J., Hsiao, H.H.: Spatiotemporal saliency detection and salient region determination for H. 264 videos. JVCIR **24**(7), 760–772 (2013)
9. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE TPAMI **35**(1), 221–231 (2013)
10. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: IEEE CVPR. pp. 1725–1732 (2014)
11. Kauffmann, J., Esders, M., Montavon, G., Samek, W., Müller, K.R.,: From Clustering to Cluster Explanations via Neural Networks. arXiv preprint arXiv:1906.07633 (2019)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Adv. in NIPS. pp. 1097–1105 (2012)
13. Landecker, W., Thomure, M.D., Bettencourt, L.M., Mitchell, M., Kenyon, G.T., Brumby, S.P.: Interpreting individual classifications of hierarchical networks. In: IEEE Symp. CIDM. pp. 32–38 (2013)
14. Lapuschkin, S., Binder, A., Montavon, G., Müller, K.R., Samek, W.: The LRP toolbox for artificial neural networks. Journal of Machine Learning Research **17**(114), 1–5 (2016)
15. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking Clever Hans predictors and assessing what machines really learn. Nature Communications **10**, 1096 (2019)
16. Li, J., Liu, Z., Zhang, X., Le Meur, O., Shen, L.: Spatiotemporal saliency detection based on superpixel-level trajectory. Signal Processing: Image Communication **38**, 100–114 (2015)
17. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
18. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1 – 15 (2018)
19. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K.: Wavenet: A generative model for raw audio. In: The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016. p. 125 (2016)
20. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: IEEE CVPR. pp. 4151–4160 (2017)
21. Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., Macdonell, C., Anvik, J.: Visual explanation of evidence with additive classifiers. In: Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA. pp. 1822–1829 (2006)
22. Reed, S.E., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., Rabinovich, A.: Training deep neural networks on noisy labels with bootstrapping. In: 3rd International

Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015)
23. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135–1144 (2016)
24. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.: Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems **28**(11), 2660–2673 (2017)
25. Schütt, K., Kindermans, P.J., Felix, H.E.S., Chmiela, S., Tkatchenko, A., Müller, K.R.: SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In: Adv. in NIPS. pp. 992–1002 (2017)
26. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE CVPR. pp. 618–626 (2017)
27. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. CoRR **abs/1511.04119** (2015)
28. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings (2014)
29. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015)
30. Srinivasan, V., Lapuschkin, S., Hellge, C., Müller, K.R., Samek, W.: Interpretable human action recognition in compressed domain. In: IEEE ICASSP. pp. 1692–1696 (2017)
31. Sturm, I., Lapuschkin, S., Samek, W., Müller, K.R.: Interpretable deep neural networks for single-trial EEG classification. Journal of Neuroscience Methods **274**, 141–145 (2016)
32. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al.: Going deeper with convolutions. In: IEEE CVPR. pp. 1–9 (2015)
33. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: IEEE ICCV. pp. 4489–4497 (2015)
34. Yang, C., Rangarajan, A., Ranka, S.: Visual explanations from deep 3d convolutional neural networks for alzheimer's disease classification. CoRR **abs/1803.02544** (2018)
35. Yosinski, J., Clune, J., Nguyen, A.M., Fuchs, T.J., Lipson, H.: Understanding neural networks through deep visualization. CoRR **abs/1506.06579** (2015)
36. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: IEEE CVPR. pp. 4694–4702 (2015)
37. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833 (2014)
38. Zhang, C., Tian, Y.: Automatic video description generation via LSTM with joint two-stream encoding. In: ICPR. pp. 2924–2929 (2016)
39. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: IEEE CVPR. pp. 2921–2929 (2016)

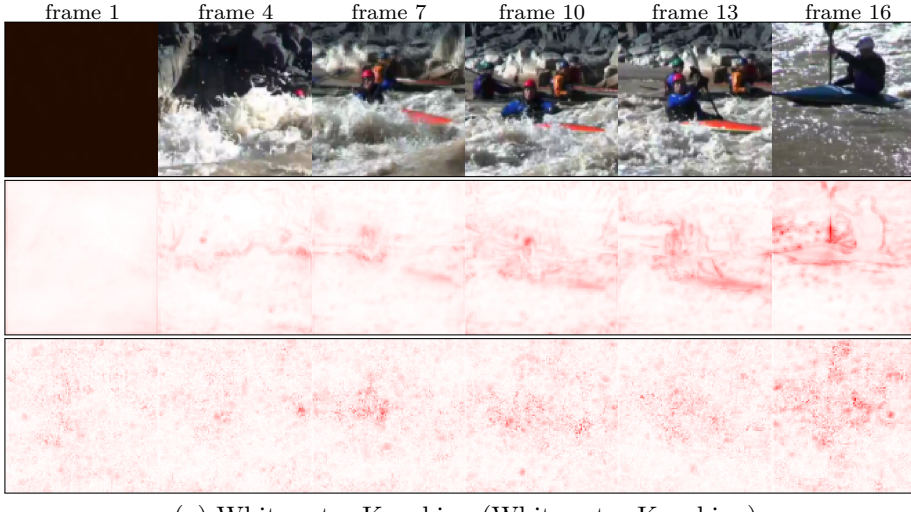


(a) Demolition Derby (Carting)



(b) Jōdō (Kenjutsu)

Fig. 16.6. Examples of videos belonging to different classes. For each example from top to bottom: input video, deep Taylor / LRP decomposition, sensitivity analysis. Captions are the true label followed by the predicted label in parentheses.



(a) Whitewater Kayaking (Whitewater Kayaking)

(b) Mushing (Gridiron Football)

Fig. 16.7. Examples of videos belonging to different classes. For each example from top to bottom: input video, deep Taylor / LRP decomposition, sensitivity analysis. Captions are the true label followed by the predicted label in parentheses.