PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Noise intensity prediction from video frames using deep convolutional neural networks

Leonardo O. MAZZA[(1)], José Gabriel R. C. GOMES [(2)], Julio Cesar B. Torres[(3)]

[(1)]Universidade Federal do Rio de Janeiro/PEE/PADS, Brazil, leonardomazza@poli.ufrj.br
[(2)]Universidade Federal do Rio de Janeiro/PEE/PADS, Brazil, gabriel@pads.ufrj.br
[(3)]Universidade Federal do Rio de Janeiro/PEE/PADS, Brazil, julio@poli.ufrj.br

Abstract

Some closed-circuit television (CCTV) systems do not have microphones. As a result, sound intensity information is not available in such systems. We present a method to generate traffic noise intensity estimates using solely video frames as input data. To that end, we trained a fully connected layer on top of VGG16 (pretrained with ImageNet) using a dataset that was automatically generated by a single camera with a mono microphone pointing at a busy traffic crossroad with cars, trucks, and motorbikes. For neural network training from that dataset, color images are used as neural network inputs, and true average noise intensities are used as neural network targets. The trained neural network successfully tracked trending noise intensities with correlation 0.594 despite their blindness to the data temporal properties. These results suggest that average noise intensity targets are sufficient for convolutional neural networks to detect noise generating sources within a traffic scene.

Keywords: Convolutional neural network, Traffic noise intensity, Non-linear regression, Non-linear prediction

1 INTRODUCTION

Noise pollution in known to have negative non-auditory effects [1]. Annoyance, health issues and learning disabilities are associated with constant exposure to high noise intensity [1]. A study [2] conducted in Berlin with 1881 patients and 2234 controls showed 1.3 times the odds of myocardial infarction in men subject to 70 dB(A) (A-weighted dB) compared to men exposed to 60 dB(A) or less. The study also demonstrated that men that lived in the area with noise intensity 70 dB(A) or higher for over 10 years presented 1.8 times the probability of myocardial infarction. A review [3] of several studies concluded that environmental noise had significant effects on academic performance of kindergartners and children at early school years. Since noise has an impact over daily life, health and intellectual development, monitoring noise intensity is a sensible action for urban environment improvement.

Aircraft, traffic and industry are usual noise sources. Traffic in particular is ubiquitous and its monitoring is already in place in the form of closed-circuit television. However some legislatures, for example California law [4], prohibits audio recordings without two party consent. As a result, several closed-circuit television cameras either do not contain a microphone or have it disabled. To be able to monitor noise when microphones are not always available we present a method to estimate noise intensity from video data (more specifically, still images without audio) only.

For that, we constructed a dataset based on a set of video sequences. This dataset contains still images (i.e. frames) and their respective average noise intensity levels. Frames are fed to a model whose parameters are optimized to minimize error between the model outputs and true noise levels. Since convolutional neural networks (CNNs) are currently the edge-performance models in image classification, segmentation, localization and other computer vision problems [5, 6, 7, 8, 9], they were selected for this task. Particularly we use the convolutional part of a pretrained model as a feature extractor, and train a fully connected (FC) network on its top. After training, the model should approximate average sound intensities from unseen frame inputs. We also train a second network to predict average sound intensity, and then use a visualization technique to evaluate whether

Figure 1. Sample frames from videos 1, 3 and 6.

the network detects noise-generating sources.

2 DATASET

Ten videos were created from a camera with a mono microphone pointing towards a busy crossroad. We downsampled the RGB videos from the original 720×480 resolution to 240×240 and extracted their audio information into wave files. From that, at each time t we define a pair of related objects: the input frame $F_t \in \mathbb{R}^{240\times240\times3}$ and the average sound intensity $S_t \in \mathbb{R}$. We define S_t as:

$$S_t = \ln\left(\frac{1}{M}\sum_{k=t-t_b}^{t+t_f} I_k^2\right), \tag{1}$$

where I_k is the value of an audio sample from the respective wave file at time k. Parameters t_b and t_f are respectively the backward and forward times considered for S_t. The number of samples I_k between time instants $(t-t_b)$ and $(t+t_f)$ is M. The camera microphone was not calibrated so noise intensity levels are not in dB.

For neural network training and validation, we selected four videos: three videos for training and one video for validation. In the training set, two videos were from a daylight traffic scene and one video was from a night-time traffic scene. Values for t_b and t_f were set to 300 ms. The validation set consisted of frames from a single night-time video. Frames were sampled once every second, which yields 1183, 1242, 1325 and 1280 frames from each video respectively. Daylight S_t sample sequences are clearly different from night-time S_t sample sequences: daylight S_t values range from 12 to 17, and night-time S_t values range from 10 to 16. In order to be applied to video sequences shot at different locations, the models trained on the previously described dataset might require the application of different output offset values. Without any further visual clue, sound intensity averages change from one particular environment to another one.

Challenges in this dataset include label noise from extraneous audio sources with no corresponding object in the video. For example, at times buses might break outside of the crossroad frame, which causes a peak in the S_t sample sequence without the presence of the bus itself in the image. Also, cars occasionally accelerate abruptly, and loud skid noise is heard without significant change in the video frame. Another issue arises from human voices from people near the microphone but not present in the screen itself. These problems were not observed to be common enough to significantly impair neural network training.

3 VGG16 AND TRAINING

The VGG16 neural network [10] is known for winning of the localization challange, and attaining second place in the classification challenge at the ImageNet contest of 2014 [11]. Its convolutional structure contains approximately 14 million parameters and, the fully connected (FC) top layers contain 124 million parameters. To build our model, we use the VGG16 convolutional part as a feature extractor by removing the top FC layer;

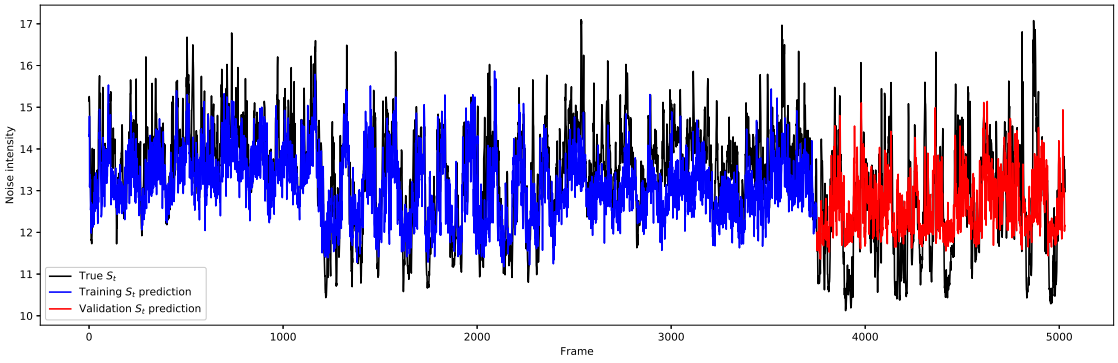

Figure 2. FC-512-128-1 with validation error 1.167 and initial learning rate: 0.01

that reduced the overall parameter count by nearly 90%. We then add a global average pooling (GAP) [12] layer at the end of the last convolutional layer, and cascade it with a two-layer FC network: the first layer has 512 inputs and 128 neurons, and the output layer has a single linear neuron. The total parameter count for these layers is 65793. We refer to this network as FC-512-128-1. Replacing the top layer for GAP+FC was shown [13] to cause small decrease in classification performance with the benefit of much better parameter efficiency. This VGG16 was pretrained with a pre-processing that subtracts channels R, G and B by 103.939, 116.779, 123.68 respectively. However we followed the pretraining described in [14]. This pretraining first computes the mean and the standard deviation for each image channel with the entire training data. Then each image channel is subtracted by its respective mean and divided by its respective standard deviation. Adam optimizer [15] was used to minimize the mean squared error (MSE) between the network output and the target S_t. For a minibatch with size N, $l(\theta) = \sum_{t=1}^{N} \frac{||f(F_t,\theta)-S_t||^2}{N}$ is the loss function to be minimized, where θ represents the FC parameters and f is the complete neural network. The convolutional layer parameters were not trained. Every single frame F_t was associated with a single S_t value, and the (F_t, S_t) pairs were randomized during training, so no temporal information is used in neural network training. FC-512-128-1 was trained for 70 epochs with initial learning rate of 0.01, which was scheduled to be multiplied by 0.3 at epochs 30 and 50. These hyperparameters were chosen from a pool of candidates. Three initial learning rates were tested: 0.1 , 0.01 and 0.001. For each of these, two optimizers were tried: Adam and SGD. For each combination the previous hyperparameters, a GAP layer was either omitted or added. And for all combinations of the mentioned, a hidden layer with 128 neurons was either omitted or added to the last FC layer. After training models with these settings, we selected those with lower validation error. For example, Figures 2 and 3 show predictions of the model FC-512-128-1 with two different settings: the first had an initial learning rate of 0.01 and the second an initial learning rate of 0.001 respectively.

3.1 Noise source detection

A second network was trained with the purpose of identifying the noise source. We refer to source detection network as FC-512-1. In a way similar to what was done for FC-512-128-1, the FC-512-1 network was also trained to predict S_t values by minimizing MSE between the network outputs and the true S_t values. Also as in the FC-512-128-1 case, the original FC layers from VGG16 network were replaced by GAP+FC. This new network was trained in the same setting as FC-512-128-1. The main difference is that the FC network added after the GAP is a single FC layer, i.e. a single neuron with 512 inputs. This allows the application of class activation maps (CAM) [13] for the visualization of image regions responsible for the prediction. In spite of this not being a classification task, the same principle holds: large network outputs should be associated with large values in the CAM. This FC network has 512 inputs and a single output, for a total of 513 added parameters.

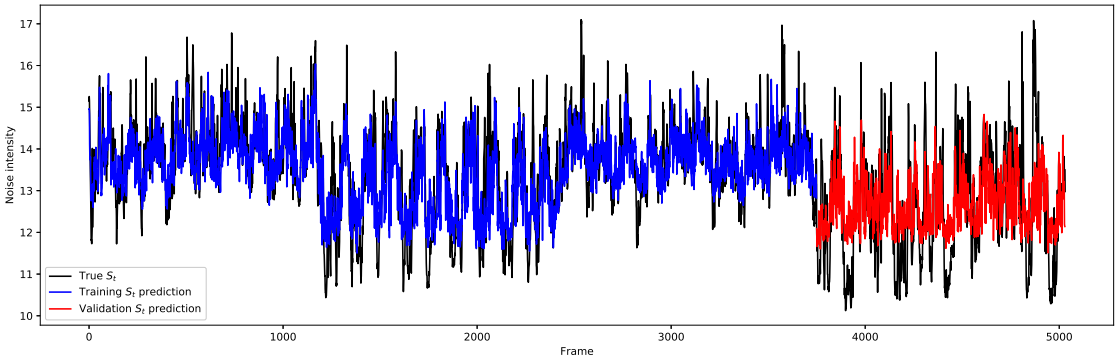

Figure 3. FC-512-128-1 with validation error 1.188 and initial learning rate: 0.001

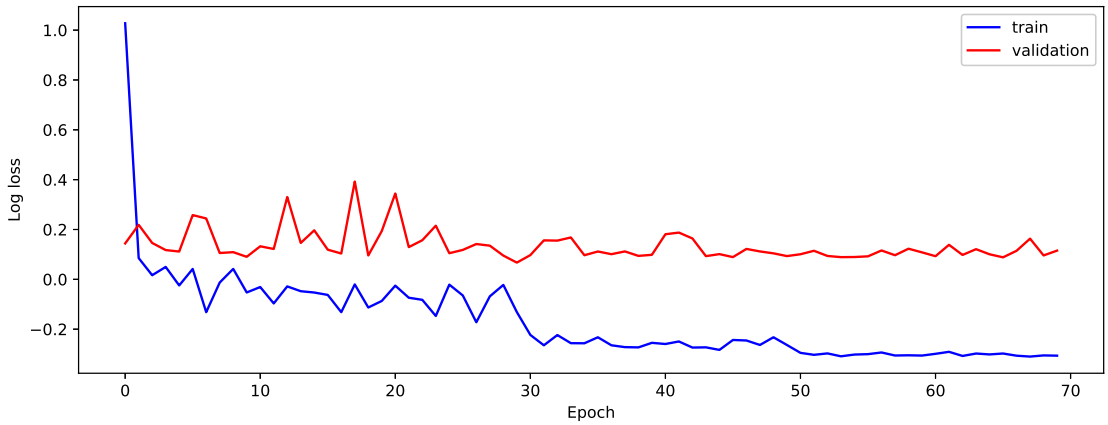

Figure 4. Training and validation FC-512-128-1 log loss curves. Validation error oscillated during the first epochs and then stalled until the first learning rate reduction at epoch 30.

The last convolutional layer of VGG16 yields 512 7×7 feature maps, which constrains object detection resolution to that 7×7 grid. To visualize the CAM, we upsample the 7×7 grid to 240x240 by nearest neighbor interpolation, and replace the green channel of the crossroad frame with the CAM. We observed that some portions of the CAM remained fixed for several consecutive frames. These fixed values had a large magnitude compared to CAM variations that define the detected object which made the changes in the green channel too subtle to be visualized. So we first subtract that average CAM (AC) offset then apply a gain α in the green channel to magnify the CAM variations. We compute an online 7×7 AC by:

$$AC[n] = AC[n-1]\lambda + C[n-1](1-\lambda) \tag{2}$$

$$DC[n] = \alpha(C[n] - AC[n]) \tag{3}$$

where $C[n] \in \mathbb{R}^{7\times7}$ is the CAM at frame n, $AC[n] \in \mathbb{R}^{7\times7}$ is the AC at frame n, $DC[n] \in \mathbb{R}^{7\times7}$ is the displayed CAM at frame n, $\lambda \in (0,1)$ and $\alpha \in \mathbb{R}$. The closer λ is to 1, the longer the effective sample sequence that is taken into account for CAM averaging. We selected $\lambda = 0.95$ and $\alpha = 10$. For noise source detection we

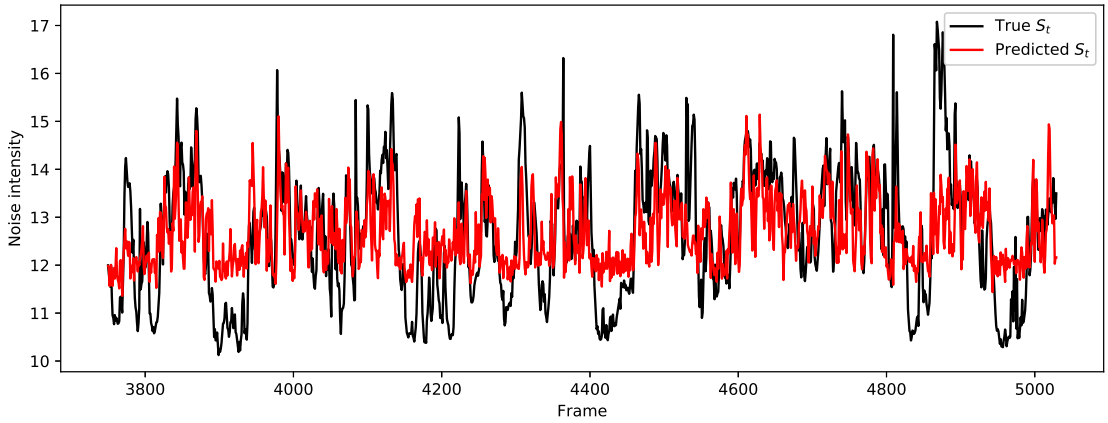

Figure 5. FC-512-128-1 validation set targets and respective outputs (predictions). The S_t values at frames 4800 and 4860 are largely missed by the network. The corresponding video shows an event in which a bus breaks, which happens outside of the camera view, creates the spike at 4800, then a motorbike skids at 4860, and that causes the second spike at that time. As expected, it was not possible to correctly infer S_t values in those cases, because the associated objects are not present in the image.

do use temporal information to filter the CAM, however this method only requires frames and no online audio data.

4 RESULTS

The training achieved a minimum validation MSE of 1.167 (Figure 4). The network predicted average noise intensity with correlation 0.594. It was capable of following trending S_t samples. The network had worse performance during low signal intensity intervals. It is possible that in spite of the significant differences between the real S_t and the network prediction during low noise intensity intervals (Figure 5, frames 3880 to 3940, among other intervals) the image had no recognizable noise generating objects throughout these intervals. Figure 5 shows that in spite of label noise, the network learned correct maps between visual objects and salient S_t values: although peaks appear in the true S_t at frame 4860, peaks are not present at the network output because of the nonexistence of potential audio sources in the image at these times.

4.1 Noise source detection

FC-512-1 achieved a minimum validation MSE of 1.181 (Figure 6). The network predicted average noise intensities with correlation 0.597. Its S_t predictions are very similar to that of FC-512-128-1 in spite of higher MSE. Figure 8 displays noise source detection results with some frames from video 8, which was used for validation. Green spots indicate pixel locations from which S_t values originate. The brighter the spot, the higher the noise the neural network predicts from that location. These results show that the network not only associates objects to S_t values to be able to predict noise intensity, but it also detects the noise source within the frame.

Overall, FC-512-128-1 and FC-512-1 achieved similar results in predicting S_t values, in spite of the lower complexity of FC-512-1. This indicates that for this number of training samples the model does not require an extra FC layer to learn the relationship between F_t and S_t. In this situation, FC-512-1 seemed a better solution to predict S_t values since it not only achieves similar performance to FC-512-128-1, but also allows visualization of the relevant image regions. Moreover, minimization of MSE for a model that is a single neuron (such as FC-512-1) is a quadratic program, a convex optimization problem with convergence guarantees.

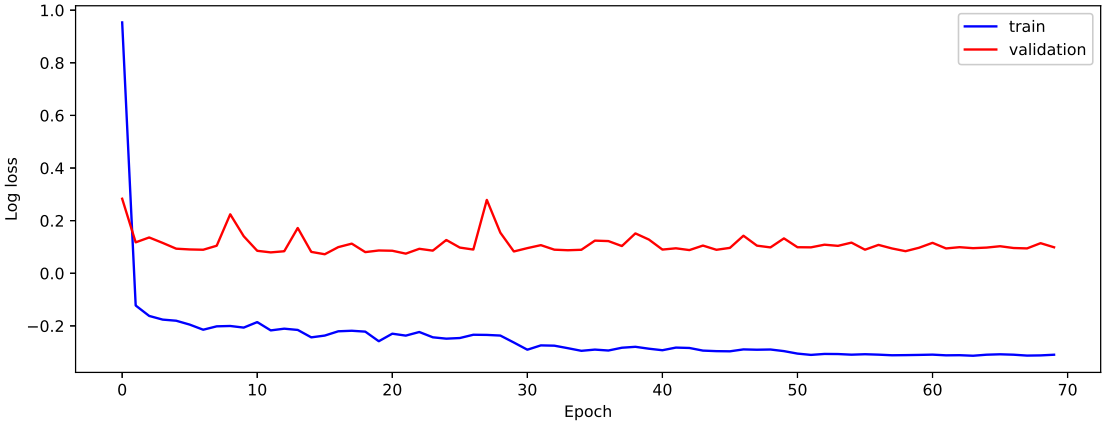

Figure 6. Training and validation FC-512-1 log loss curves. The validation loss decreases and stalls after the initial epochs.

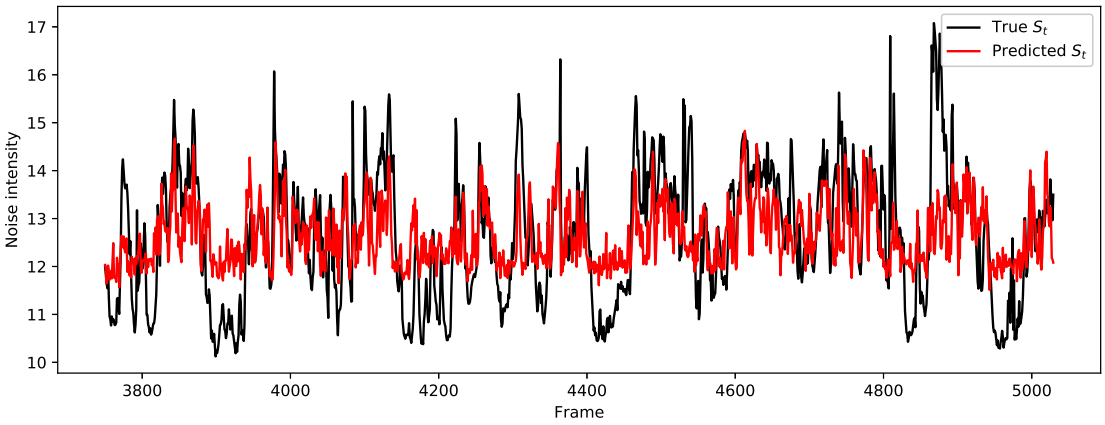

Figure 7. Network outputs for the validation set of FC-512-1. Results are similar to those in Figure 4 in spite of its lower complexity and lower parameter count. Slight differences can be seen at Frame 4120.

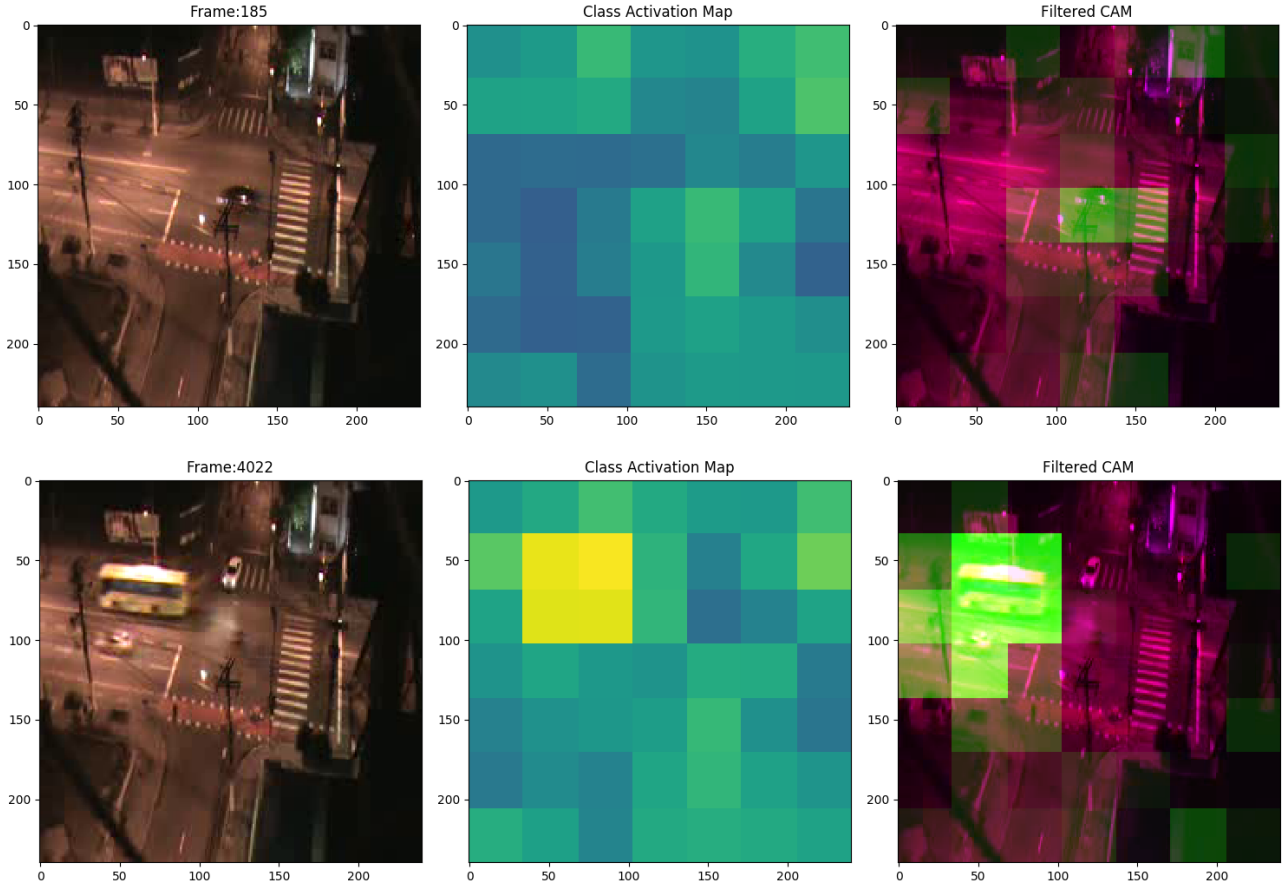

Figure 8. Left: original frame. Middle: unprocessed CAM. Right: original frame with DC in the green channel. These frames are from video 8, and were not used for training. The top row shows a typical example: a car drives though the crossroad with some unwanted spots around the image. The bottom row shows a particularly good result: when a noisy bus drives through the crossroad.

5 CONCLUSION

FC-512-128-1 and FC-512-1 were able to follow trends in average noise intensity levels. This occurred in spite of challenges in the dataset such as occasionally high S_t values without corresponding noise sources within the frame. We compared the performance of the two networks and concluded that both predicted S_t values similarly. FC-512-1 in particular allowed the use of CAM to visualize image regions responsible for network ouputs. We had to temporally filter the CAM to be able to observe changes that detect the noise sources. We did that by computing a online average CAM that was subtracted from the unprocessed CAM. This result was multiplied by a gain, upsampled to 240×240 by nearest neighbor interpolation and set as the green channel of the video frame. That produced a visualization where bright green spots represent the detection of a source. Finally, that helped us verify that the network predicts noise intensity by associating objects to S_t values.

ACKNOWLEDGEMENTS

The authors thank the financial support of CAPES/DAAD PROBRAL Program for developing this work through project number 88881.198848/2018-01, and the financial support of CNPq through projects 432997/2018-0 and 309602/2016-5.

REFERENCES

[1] Stephen A Stansfeld and Mark P Matheson. Noise pollution: non-auditory effects on health. *British Medical Bulletin*, 68(1):243–257, December 2003. doi: 10.1093/bmb/ldg033. URL https://doi.org/10.1093/bmb/ldg033.

[2] Mette Sørensen, Zorana J. Andersen, Rikke B. Nordsborg, Steen S. Jensen, Kenneth G. Lillelund, Rob Beelen, Erik B. Schmidt, Anne Tjønneland, Kim Overvad, and Ole Raaschou-Nielsen. Road traffic noise and incident myocardial infarction: A prospective cohort study. *PLoS ONE*, 7(6):e39283, June 2012. doi: 10.1371/journal.pone.0039283. URL https://doi.org/10.1371/journal.pone.0039283.

[3] Gary W. Evans and Stephen Lepore. Nonauditory effects of noise on children: A critical review. *Children's Environments*, 10:31–51, 01 1993. doi: 10.2307/41515250.

[4] California penal code paragraph 632, January 2017. URL https://leginfo.legislature.ca.gov/faces/codes_displaySection.xhtml?lawCode=PEN&sectionNum=632. Accessed 23 April 2019.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012. doi: 10.1145/3065386.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. pages 770–778, 06 2016. doi: 10.1109/CVPR.2016.90.

[7] Ross Girshick Jian Sun Shaoqing Ren, Kaiming He. Faster R-CNN: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[8] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[11] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[12] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network, 2013.

[13] B. Zhou, A. Khosla, Lapedriza. A., A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *CVPR*, 2016.

[14] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2016.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.