TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Khademi, S., Shi, X., Mager, T., Siebes, R., Hein, C., De Boer, V., & Van Gemert, J. (2018). Sight-seeing in the eyes of deep neural networks. In W. Hazeleger (Ed.), *Proceedings - IEEE 14th International Conference on eScience, e-Science 2018* (pp. 407-408). Article 8588744 IEEE. https://doi.org/10.1109/eScience.2018.00125

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' – Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

2018 IEEE 14th International Conference on e-Science

Sight-Seeing in the Eyes of Deep Neural Networks

Seyran Khademi*, Xiangwei Shi *, Tino Mager†, Ronald Siebes ‡, Carola Hein †, Victor de Boer ‡ and Jan van Gemert *

*TU Delft, Faculty of Electrical Engineering, Mathematics and Computer Science, Department of Intelligent Systems
†TU Delft, Faculty of Architecture and the Built Environment, Department of Architecture
‡ VU Universiteit Amsterdam Faculty of Sciences, Department of Computer Science

***Abstract*—We address the interpretability of convolutional neural networks (CNNs) for predicting a geo-location from an image. In a pilot experiment we classify images of Pittsburgh vs Tokyo and visualize the learned CNN filters. We found that varying the CNN architecture leads to variating in the visualized filters. This calls for further investigation of the effective parameters on the interpretability of CNNs.**

***Index Terms*—convolutional neural network (CNN), interpretability, place recognition, visualization, classification.**

I. Context

We investigate what visual cues can discriminate visual geo-locations. We draw inspiration of [1], however using modern deep learning methods to learn discriminative features in city views. These features can be exploited by researchers in the humanities to study various aspects of urban and architecture design as well as its social attributes.

Human interpretability of intelligent systems is a key factor for establishing trust between the user and the machine [2]. Initial attempts to visualize the learned attributes in convolutional neural networks have commenced since the advent of CNN to unfold the magic of the black box [3]–[5]. There is yet an increasing interest in probing these popular deep neural networks (DNN) [6]–[9]. We track the emerge of semantic objects at the final layer representation of CNN as in [9].

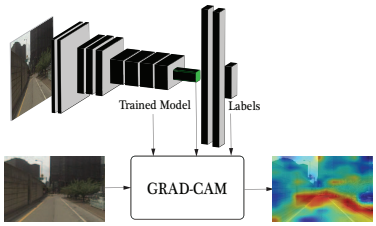


Fig. 1. Grad-Cam [10] visualizes a trained CNN model using the ground truth label and the test image. The output is a corresponding importance heat-map showing the most and the least discriminative areas with red (high value) to blue (low value) colors, respectively.

II. Method & Results

We use the recent Grad-Cam [10] to investigate how CNN architectures vary in their interpretability (Fig. 1). We consider three models in a visual place recognition (classification) task between images of Tokyo and images of Pittsburgh [11]: 1. a shallow (four convolutional layers and two fully connected layers with max pooling and ReLu activation layers in between), 2. the VGG11 model [12] and 3. the ResNet18 model [13]. All three models are trained using the cross-entropy loss. The training, validation and test datasets are constructed with the proportions as 6:2:2, respectively. Training sets are balanced and consists of 45,000+ samples.

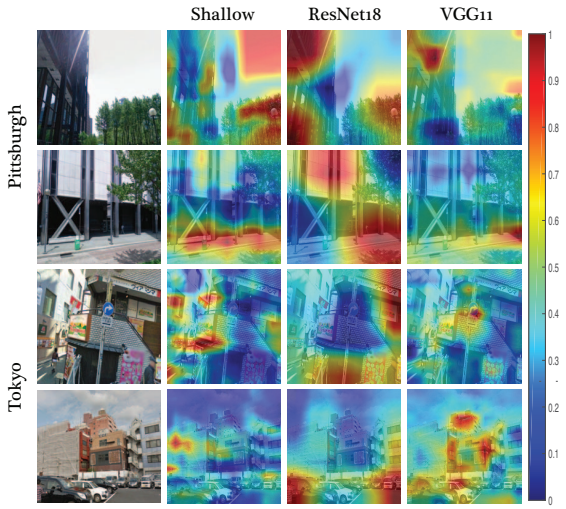


Fig. 2. Different CNN models learn dissimilar attributes for place recognition. Note that a shallow net triggers on the sky or on disjoint regions in the image. The ResNet focuses on wider regions and VGG is more selective.

For all three models the test set classification accuracy is consistently over 99%. The visualizations (Fig. 2), however, show high variation between networks. Our observations indicate that VGG11 shows more semantically meaningful representation at the final convolutional layer compared to the ResNet18 and the shallow CNN. Moreover, the shallow CNN picks up on the unwanted bias in the datasets, e.g. a clear or cloudy sky than the deeper CNN models. Finally, VGG11 most often highlights pathways for Pittsburgh , while in Tokyo it selects kanji signs as the most discriminative attributes.

References

[1] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros, "What makes paris look like paris?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 101:1–101:9, Jul. 2012.

978-1-5386-9156-4/18/$31.00 ©2018 IEEE
DOI 10.1109/eScience.2018.00125

IEEE computer society

Authorized licensed use limited to: TU Delft Library. Downloaded on May 07,2020 at 09:00:14 UTC from IEEE Xplore. Restrictions apply.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "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*, 2016, pp. 1135–1144.
[3] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, vol. abs/1311.2901, 2013.
[4] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *CoRR*, vol. abs/1312.6034, 2013.
[5] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller, "Striving for simplicity: The all convolutional net," *CoRR*, vol. abs/1412.6806, 2014.
[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene CNNs," in *International Conference on Learning Representations (ICLR)*, 2015.
[7] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," *CoRR*, vol. abs/1704.05796, 2017.
[8] R. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," *CoRR*, vol. abs/1704.03296, 2017.
[9] A. Gonzalez-Garcia, D. Modolo, and V. Ferrari, "Do semantic parts emerge in convolutional neural networks?" *Int. J. Comput. Vision*, vol. 126, no. 5, pp. 476–494, May 2018.
[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 618–626.
[11] R. Arandjelovic, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: CNN architecture for weakly supervised place recognition," *CoRR*, vol. abs/1511.07247, 2015.
[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

Authorized licensed use limited to: TU Delft Library. Downloaded on May 07,2020 at 09:00:14 UTC from IEEE Xplore. Restrictions apply.