DEEPLING: A VISUAL INTERPRETABILITY SYSTEM FOR CONVOLUTIONAL NEURAL NETWORKS

Daniel Weidele [1] **Hendrik Strobelt** [1] **Mauro Martino** [1]

ABSTRACT

We demonstrate an interactive visualization system to promote interpretability of convolutional neural networks (CNNs). Interpretation of deep learning models acts on the interface between increasingly complex model architectures and model architects, to provide an understanding of how a model operates, where it fails, or why it succeeds. Based on preliminary expert interviews and a careful literature review we design the system to comprehensively support architects on 4 visual dimensions.

1 OVERVIEW

Building artificial intelligence systems involves a good degree of trial and error in establishing an appropriate architecture and hyperparameter settings, and large to massive amounts of training data. The operation of these systems, once constructed can be mysterious to their developers, and their function can be regarded as a black box. Consequently, while their accuracy can be tested, the nature, cause, and potential resolution of the errors that they make is not visible to the developer, and thus changes to improve performance are difficult to predict.

Convolutional neural network (CNN) architectures were first explored visually by Zeiler & Fergus (2014). Harley (2015) visualizes CNNs and their activation patterns threedimensionally. While there have been developed several fine-grained visualization techniques and extensions of such, it is only recently that more holistic *visualization systems* are being proposed. Yosinski et al. (2015) is probably the first attempt in a series, offering a comprehensible graphical user interface to explore activations, gradients, or deconvolutions. In CNNVis Liu et al. (2017) further hybridize visualization techniques to enable visual analytics on CNNs–while not in their focus, also simultaneously incorporating a basic representation of the model graph itself.

We here propose the demonstration of a visual interpretability system, that is first and foremost driven by preliminary expert interviews and a careful literature review. We identify a chance to comprehensively support architects on 4 visual dimensions: architecture centric, data centric, model semantic and via What If? scenarios. Especially when put into action together these valuably add to the model architect's toolbox. We demonstrate our resulting implementation of the system that runs with minimal setup requirements for users. The software launches directly on top of trained CNNs with injected user data.

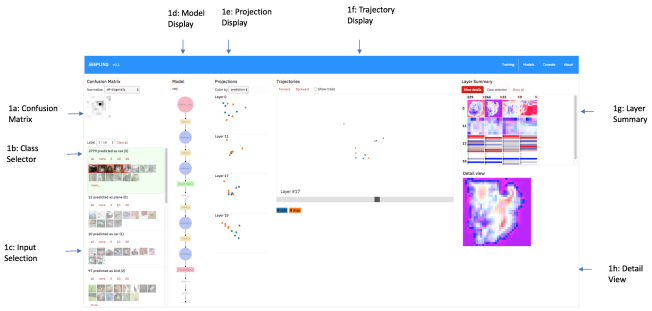

Figure 1. Overview of the system "Deepling", subject of the demonstration.

2 DEMONSTRATION

We demonstrate our reference implementation of the visual interpretability system (Fig. 1) live and interactively to a small audience. The demonstration is scheduled as blocks of 10 minutes, running 6 times per hour of demonstration. In the first 5 minutes the presenter will outline the ideas behind the software by directly guiding through the graphical user interface, hosted on an AlexNet (Krizhevsky et al., 2012) implementation and CIFAR-10[1] data set. We additionally outline how users can connect their own convolutional neural network implementations to the system. The latter 5

[1]MIT-IBM Watson AI Lab, IBM Research, Cambridge (MA), USA. Correspondence to: Daniel Weidele <daniel.weidele@gmail.com>.

Proceedings of the 2^{nd} SysML Conference, Palo Alto, CA, USA, 2019.

[1] www.cs.toronto.edu/~kriz/cifar.html

minutes are meant to address questions of spectators, and allow single individuals to directly interact with our system.

3 Elements of Novelty

3.1 Adversarial strips

We propose a novel visualization to get an understanding of a neurons meaning. Like Mordvintsev et al. (2015) we modify the input image to optimize a desired activation pattern. However, by showing the differences gradually we can more easily perceive the emerging pattern, and which parts of the input image are salient to change. Figure 2 shows a resulting explanation for a plane being confused with a horse by the CNN model.

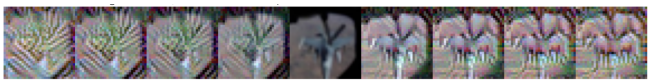

Figure 2. Adversarial strip for an input picture of ground truth "airplane". The CNN model classified the picture as "horse". The adversarial strip increasingly interferes the model's understanding of "horse" with the input picture, to help us understand the driving factors for misclassification: we may assume planes are usually not depicted from a top-down view - so jet engines pointing downwards may have been confused with horse feet.

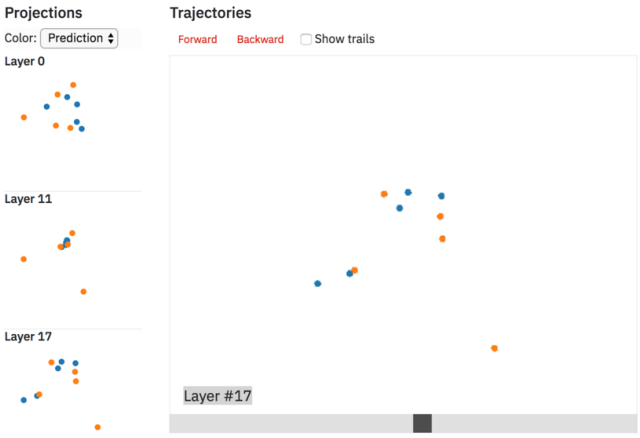

Figure 3. Projections of 5 images of class cat (orange), and 5 images of class dog (blue) across layers 0, 11, and 17. The animation view in the center transitions between the layers interactively.

3.2 Animation loops between layer-wise projections of neuron acticatios

We project activation patterns of user-selected data per layer via PivotMDS (Brandes & Pich, 2006). To align the projections, we use Prokrustes analysis with additional test for mirroring. Transitions between the obtained layouts are animated using a custom WebGL component. The result is an impressive macro-scale insight into a CNN's operation. Figure 3 depicts the graphical interface view within which projections are animated. The user can interact with the slider control at the bottom to seemlessly animate between layer projections.

4 Equipment

4.1 Equipment brought by the demonstrator

The demonstration is performed on a laptop by the presenter. The presenter will bring their own laptop, as well as a charging cable.

4.2 Requirements at the place of demonstration

We have simple requirements:

- Bar table
- Power socket in reach
- Monitor with HDMI plug for laptop

References

Brandes, U. and Pich, C. Eigensolver methods for progressive multidimensional scaling of large data. In *International Symposium on Graph Drawing*, pp. 42–53. Springer, 2006.

Harley, A. W. An interactive node-link visualization of convolutional neural networks. In *International Symposium on Visual Computing*, pp. 867–877. Springer, 2015.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Liu, M., Shi, J., Li, Z., Li, C., Zhu, J., and Liu, S. Towards better analysis of deep convolutional neural networks. *IEEE transactions on visualization and computer graphics*, 23(1):91–100, 2017.

Mordvintsev, A., Olah, C., and Tyka, M. Inceptionism: Going deeper into neural networks. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html, 2015.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.