Transparent Deep Learning

Roger Milroy
Supervised By: Raja Naeem Akram, Konstantinos Markantonakis

Information Security Group, Smart Card and IoT Security Center
Royal Holloway, University of London

The Smart Card and Internet of Things Security Centre

Objectives

To create a tool that allows us to explain in a human understandable way the decisions taken by deep neural networks.

What is Deep Learning?

- A subset of Machine learning that is characterised by deep neural networks.
- Different from older neural network techniques in that the networks learn their own features rather than features being created or selected by experts.
- Deepness is the other key difference, hence the name. Deep neural networks consisting of many layers have proven to be more successful at function approximation in highly non-linear input domains.

Why Deep Learning?

Deep learning has been extremely successful, improving performance on computer vision tasks such as image classification to super-human levels on the ImageNet database.

Unfortunately despite their high levels of performance our level of understanding of how these networks achieve it is still limited. More than that we often struggle to explain why they fail when they do, for example with adversarial examples.

AI techniques and Deep Learning in particular have already become standard in some industries and are rapidly being adopted in almost all areas of business.

Almost all mobile phones use neural networks for facial and object recognition in images and while taking pictures. But the areas where these techniques are moving into are much more sensitive including criminal justice and finance.

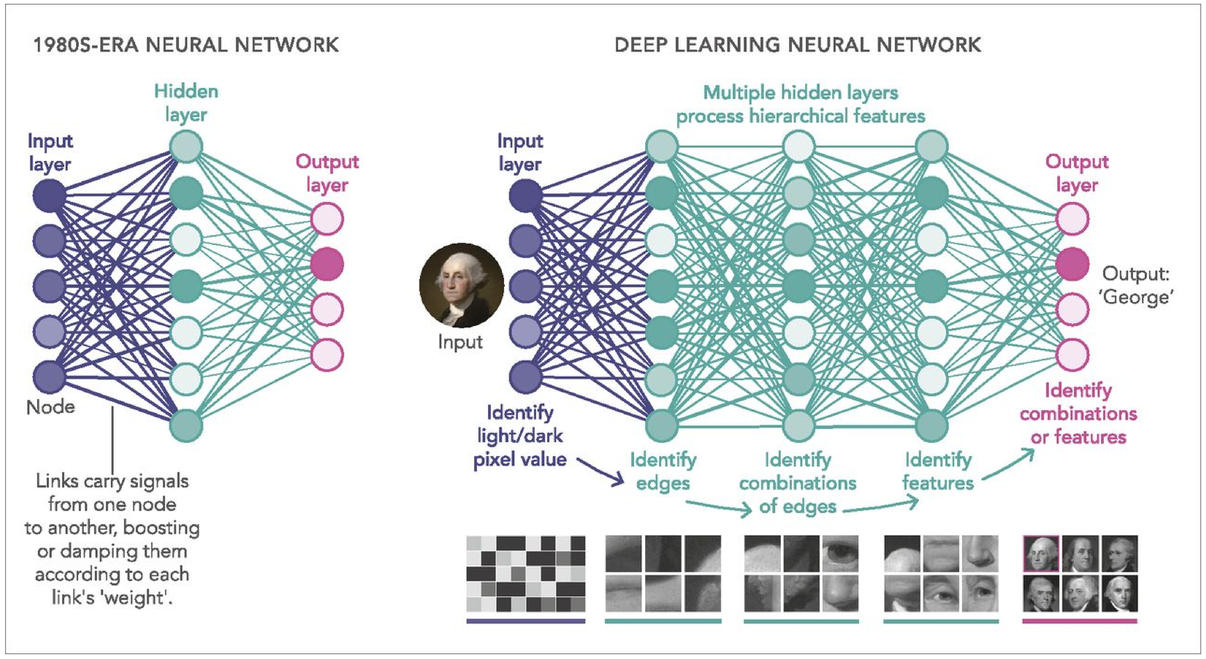

Figure 1: Deep Learning vs. Traditional Neural Networks. Image credit: Lucy Reading-Ikkanda (artist)

Explainability and Interpretability

The terms Explainability and Interpretability are related concepts that are used somewhat interchangeably.

Interpretability is essentially how easy to interpret the model or the decisions made by the model are. This usually is done by experts, either domain experts or data scientists.

Explainability is slightly different in that it is focuses on whether a models decisions can be explained. It is therefore much more focused on a single decision.

There has been a lot of work recently in both areas, one example is DeepDream [1] which focuses on interpreting what neurons in a deep neural network have learned through visualising maximal activations.

Challenges

There are a number of complications to the effort to create generalised explanation mechanisms. These can be broadly (though not exhaustively) categorised into the following areas of challenge:

- Architecture: Fully connected, Convolutional, Recurrent, ResNet, ELM and more
- Activations: Sigmoid function, Tanh, ReLU, LeakyReLU, PReLU etc.
- The number of parameters in each model. This can often run into the millions.
- The representation of the domain. Explanations of image based networks are easier for us to comprehend but it is much less intuitive to explain a network working on a high-dimensional non image input. What is the right representation?

Current Efforts

Some of the best efforts currently for explaining a decision made by a Deep Neural Network are Layer-wise Relevance Propagation (LRP) and Sensitivity Analysis (SA).

- **Sensitivity Analysis** [2] is a technique that visualises the sensitivity of the network to changes in its input. It shows which inputs need to be changed and in which direction in order to change the decision of the network.
- **Layerwise Relevance Propagation** [3] is a different technique which attempts to quantify the contribution of each input pixel to the final decision.

We have built upon these techniques while adding contextual information from training data in order to create more understandable and complete explanations.

References

[1] Alexander Mordvintsev, Christopher Olah, and Mike Tyka.
Inceptionism: Going deeper into neural networks, Jun 2015.

[2] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller.
How to explain individual classification decisions.
J. Mach. Learn. Res., 11:1803–1831, August 2010.

[3] Sebastian Bach, Alexander Binder, GrÃlgoire Montavon, Frederick Klauschen, Klaus-Robert MÃijller, and Wojciech Samek.
On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.
PLOS ONE, 10(7):1–46, 07 2015.

Contact Information

- Web: https://scc.rhul.ac.uk/
- Email: roger.milroy.2016@live.rhul.ac.uk