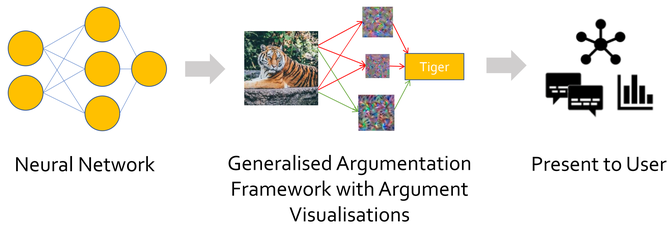
Argflow: A Toolkit for Deep Argumentative Explanations for Neural Networks

Adam Dejl, Peter He, Pranav Mangal, Hasan Mohsin, Bogdan Surdu, Eduard Voinea, Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, Francesca Toni
Department of Computing, Imperial College London

Summary

Neural network (NN) models are often difficult for human users to understand. We present **Argflow**, a toolkit enabling the generation of a variety of explanations for NN outputs in a classification setting using an argumentation-based approach called **Deep Argumentative Explanation (DAXs)**.

Deep Argumentative Explanations



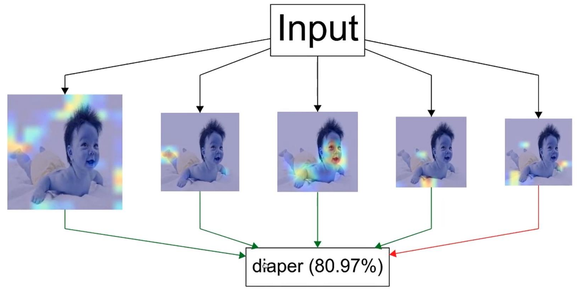
Neural Network | Generalised Argumentation Framework with Argument Visualisations | Present to User

Given a neural network $\mathcal{N}$, we can generate a directed graph $\langle N, I \rangle$ which corresponds how each neuron in $\mathcal{N}$ affects the output of other neurons in $\mathcal{N}$. In practice, N may be a subset of neurons from $\mathcal{N}$. From this, we can derive a **Generalised Argumentation Framework (GAF)** by mapping each node to an argument with some strength and each edge to a type of argumentative relation (e.g. attack). Finally, we map different arguments to human-friendly visualisations using a function χ and present our GAF to users in some modality ϕ. We refer to explanations created with this method as DAXs. A fuller treatment of DAXs can be found in [1].

[1] Emanuele Albini, Piyawat Lertvittayakumjorn, Antonio Rago, and Francesca Toni. 2020. DAX: Deep Argumentative eXplanation for Neural Networks. CoRRabs/2012.05766 (2020). https://arxiv.org/abs/2012.0576

[2] Ramprassath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. IEEE International Conference on Computer Vision (ICCV). 618–626. https://doi.org/10.1109/ICCV.2017.74

Example Explanations for VGG16



Baby or Diaper?

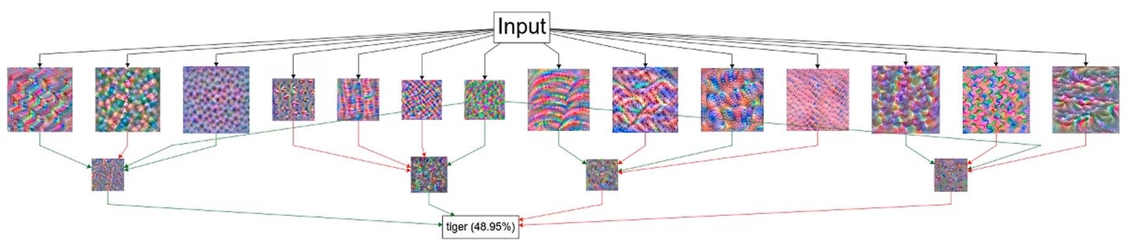
We visualise the filters in the final convolutional layer of the network using the Grad-CAM algorithm [2] as our χ and a graph visualisation as our ϕ. We can see the filters picking up parts of the baby when coming to the predicted classification 'diaper'.

Explanation Chatbots

Using the same layer and χ as before, we generate an explanation for the network's predicted classification of 'tiger', though this time picking a conversational interface as our ϕ.

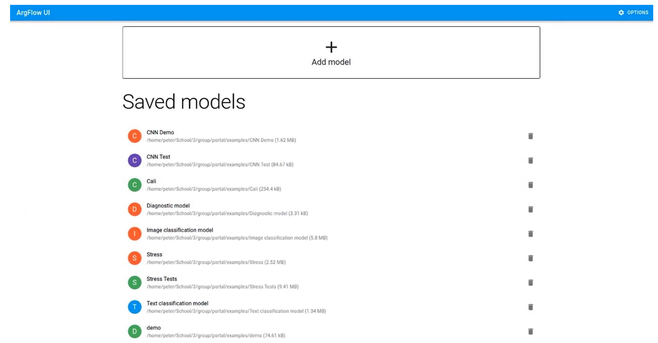
Going Deeper

We produce a deeper visualisation of the network by visualising filters from the tenth and final convolutional layers. We use activation maximisation as our χ and a graph visualisation as our ϕ.



Library

Argflow comes in the form of a Python library and a web portal implemented using Python and React. The library computes explanations and sends them to a locally-running instance of the portal for display to end users. Both the library and portal are modular and highly extensible.



Screenshot of the portal's homepage. Users can select which model they'd like to view explanations for.

Open Questions

- What are the best choices of χ, ϕ and the other mappings?
- How might we best integrate these DAXs into real-world applications?
- How might we extend DAXs to recurrent neural networks or transformers?

Links

Code: https://gitlab.com/argflow
Demo video: https://youtu.be/LPz4QbmLaxs
Correspondence: ph1718@imperial.ac.uk

AAMAS2021 LONDON, UK

Imperial College London