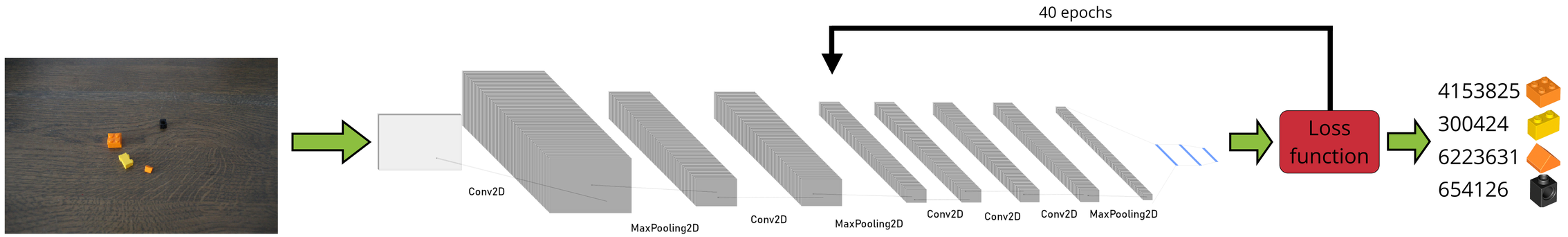
Evaluating the performance of the LIME and Grad-CAM explanation methods on a LEGO multi-label image classification task

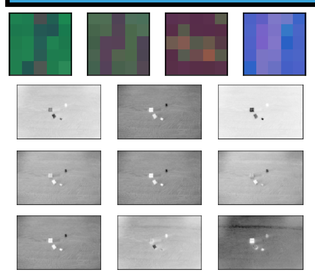
Cian, David / van Gemert, Jan (supervisor) / Lengyel, Attila (supervisor) / Delft University of Technology

Here's a question children often ask themselves: after having having dumped my entire LEGO collection on the floor, do I have the necessary bricks to build that Starwars AT-AT I'm dreaming of? Most children aren't exactly in a position to develop a deep learning model to solve this problem, which is why we did it for them. Our model takes in images of multiple LEGO bricks in a random configuration and labels them with the visible bricks.



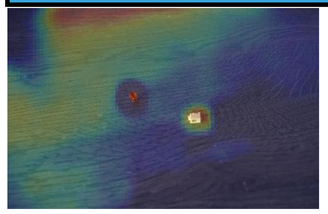
What has my network learned? How does it work? → ? → Two relevant criteria for LEGO identification model: core performance and trust

Kernel and feature map visualizations



- Visualizing convolution kernels and resulting feature maps is a rudimentary form of explanation
- Network learned noisy edge detectors
- After the first convolutional layer, impractical to visualize kernels, as they have hundreds of channels
- The deeper the convolutional layer, the more specific the response to certain bricks, as seen on feature maps

Gradient-weighted class activation mapping

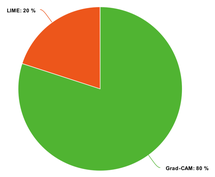


- Grad-CAM computes the sum of all the feature maps of the deepest convolutional layer, weighing each by their contribution to a prediction
- Class-discriminative: if the network distinguishes between bricks, the heatmap highlights only the right brick
- Coarse: doesn't show individual pixels, but can be used with guided backpropagation for pixel precision

Local interpretable model-agnostic explanations

- LIME creates a linear proxy model behaving similarly to real model in proximity of selected input
- Linear model takes superpixels as input: regions of similar pixels
- In theory, LIME-generated model is directly interpretable since linear
- On our lightly-trained model, performs poorly, highlighted regions hard to interpret (green is positive contribution)

Trust: which inspires the most, between Grad-CAM and LIME?



- Binary forced choice: human-grounded evaluation metric for explanation methods, where participants must choose between explanations produced by two methods
- Blind study conducted on 10 participants: they didn't know which method was which, had to choose between Grad-CAM and LIME produced explanations for 10 different input images
- Participants with all levels of familiarity with deep learning participated in study
- Result: 80% prefer Grad-CAM, 20% LIME
- Grad-CAM inspires more trust in this scenario: keep in mind, this is not a general rule

Discussion: conclusion and limits

- LIME didn't perform as well as might have been expected, but all methods have their pros and cons: we recommend using the right method for the job. Kernel and feature map visualizations are great for quick analysis, as a smoke test, and require a bit of deep learning expertise to interpret. Grad-CAM is a slower to compute, but provides meaningful, class-discriminative heatmaps. LIME offers the benefit of building a directly interpretable proxy model, but we do not recommend it until later phases of training.
- For the LEGO identification problem, using the three methods together offers considerable benefits over limiting oneself to only one
- Only two state-of-the-art methods were tried. Other promising methods include: occlusion analysis, activation maximization, integrated gradients, etc.
- The evaluation and comparison was rather informal. This is enough of a problem as is in XAI literature, and the author acknowledges the benefits a more rigorous and quantitative approach would have yielded