Shapley Values in CNN: Understanding Model Focus in Image Classification

University of South Wales

Module: MSc AI - Machine Learning and Autonomous Systems
Project: Poster Presentation

Authors
Mariusz Soltycz
30030295

1. Introduction:

What are Shapley's Values?
Shapley Values, originally from cooperative game theory [1], are used in Explainable AI (XAI) to attribute importance to input features in a model's prediction [2].

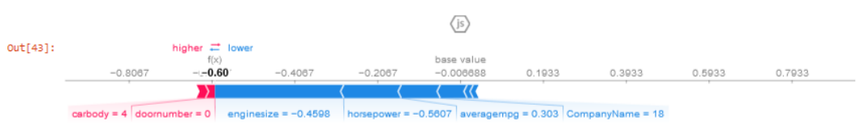
Why are they important?
In deep learning, models like CNNs achieve high accuracy, but it is unclear **what features drive their decisions**. Shapley Values help visualise which parts of an image influence the model's classification.

Comparison with Heatmap Correlation:

- **Heatmaps** show global correlations but **do not explain individual predictions**.
- **Shapley Values** reveal the specific importance of each pixel for a particular classification.

Example: Shapley Values in a Simple Model

Figure 1 Shapley Values Plot



This plot shows how different features contribute to a model's prediction. Blue values decrease confidence, while red values increase. Summing these influences results in the final prediction.

2. Research Question:

- How do Shapley Values help interpret complex CNN models analysing images?
- Do different CNN architectures change which features the model focuses on?
- Can SHAP help reveal whether the model is looking at the correct parts of an image?

3. Methodology:

Dataset: Cats vs Dogs

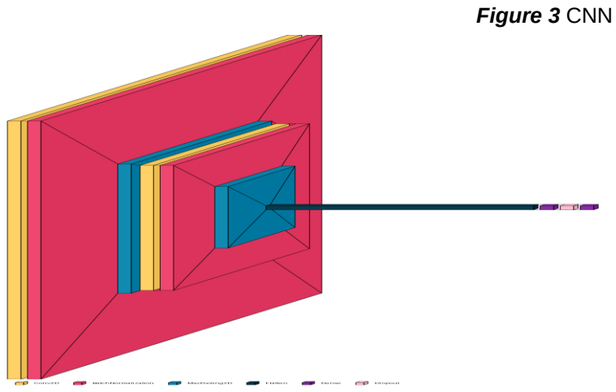
- Standard dataset from TensorFlow Datasets.
- Binary classification task (cat vs dog)

Figure 2 Sample images from Dataset

4. CNN Models Architectures:

- Basic CNN - Standard convolutional model
- Optimized CNN - More layers, batch normalization, and improved dropout.
- Augmented CNN - Data augmentation applied for robustness

Figure 3 CNN Model Structure Layers



Visualkeas[3], a Python[4] library designed to generate graphical representations of neural networks defined in Keras[5].
It creates clear visualisations that facilitate understanding the model's structure.

CNN Model Predictions:
Below are model predictions, showing some correct classifications and misclassifications.

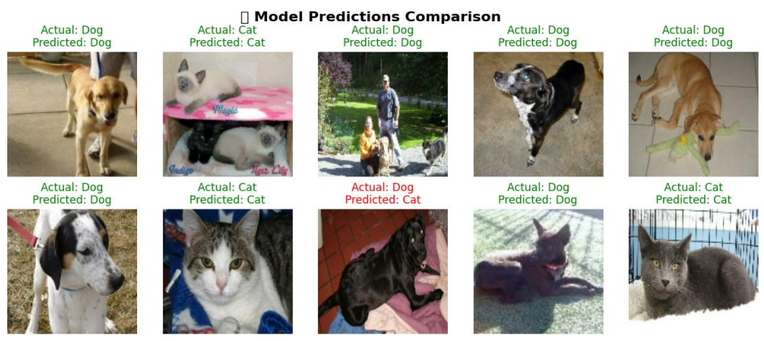


Figure 4 Basic CNN Predictions

Evaluation:
Using SHAP Heatmaps to visualize which image regions contribute most to predictions.

5. Results and Discussion:

SHAP Heatmaps :

Initial analysis shows that the model sometimes focuses on irrelevant background details rather than animal-specific features

Comparison of Models:

Figure 5 SHAP Heatmap Comparison - Different Models

Basic CNN:

- Inconsistent focus. In some images, SHAP heatmaps focus on background elements rather than the main object (dog/cat).
- **Red regions (positive influence)** often appear on **irrelevant details**, such as shadows, **background textures**, or **human presence** in the image.
- The model sometimes misclassifies dogs as cats due to their focus on non-discriminative features.

Optimised CNN:

- More stable feature recognition compared to the Basic CNN.
- SHAP heatmaps indicate improved focus on essential features, such as the shape of the animal's head, eyes, and fur texture.
- Less reliance on background information, though some noise still exist.
- Misclassifications are reduced, but some errors still occur in challenging cases, such as small dogs are confused with cats due to posture or lighting.

Augmented CNN:

- Best generalisation among all models.
- SHAP heatmaps indicate a wider distribution of feature importance, meaning the model relies on multiple image areas rather than specific spots.
- Some over-adjustment issues: The model occasionally relies too much on texture or specific details, leading to incorrect predictions when images differ slightly from training data.

Effect of Augmentation & Smoothing SHAP

- Applying SHAP smoothing improved the visualisation of critical regions.

Figure 6 Basic CNN (Smoothed SHAP)

- Augmented CNN models shift their focus compared to unoptimised ones.
- **Key finding:** Even if accuracy is high, models might base predictions on the wrong image regions.

6. Conclusion:

Shapley Values provide insights beyond accuracy, showing how CNNs interpret images. Different architectures influence what the model "sees." SHAP helps evaluate model reliability, especially in sensitive fields like medicine and finance.
Future work: Extending SHAP analysis to more complex datasets and architectures.

7. References:

[1] L. Štrumbelj, I. Kononenko (2010) "An Efficient Explanation of Individual Classifications using Game Theory," Journal of Machine Learning

[2] S. Lundberg, S.-I. Lee (2017) "A Unified Approach to Interpretable Machine Learning," NeurIPS

[3] Garcia, P. and contributors (2021) 'Visualkeras', *PyPI*. Available at: https://pypi.org/project/visualkeras/

[4] Python Software Foundation (2023) 'Python Language Reference', Python 3.10.4 documentation.

[5] Chollet, F. and others (2023) 'Keras: The Python Deep Learning API', Keras documentation.

Code Link: https://colab.research.google.com/drive/1c8MIdbQLKMC7u64rbs1pNiar7FdOdBA6?usp=drive_link