A Project report on

Interpretable Convolutional Neural Networks

By Group 15

Adwait Kulkarni - 2019A7PS0120G
Aaranya Prasad - 2019A7PS0107G*
Manthan Asher - 2019A7PS0144G
Dhyana Chidvilas Rottela - 2019A7PS0093G

Submitted in partial fulfilment of the requirements of

CS F425 - DEEP LEARNING

Under the Supervision of Prof. Tirthraj Dash

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI - K.K. BIRLA GOA CAMPUS

January 2022 - May 2022

Acknowledgement

We express our deep sense of regards and indebtedness to Professor Tirthraj Dash for his valuable guidance, continuous encouragement and wholehearted support, which were of immense help for us in completing this project titled “Interpretable Convolutional Neural Networks”. Prof. Tirthraj has been an excellent advisor, teaching us how to choose the problems to work on, how to think about the approach, and how to present them. Without his support and encouragement, this project could not have been completed.

Introduction

Neural Networks have been subject to being called “Black-boxes”. In simple terms, we know they are accurate, but we don’t know how/why. This makes it harder for industries which use these networks in real-life to trust the decisions/predictions made by it. The explicit knowledge representation in an interpretable CNN can help people understand the logic inside a CNN, i.e. what patterns are memorised by the CNN for prediction. Experiments have shown that filters in an interpretable CNN are more semantically meaningful than those in a traditional CNN. This helps resolve the dataset bias and representation bias, which may not have been detected by the test data/error.

Notebook: Colab
Github: Interpretable CNN
Weights: DL_weights

Dataset

The chosen dataset is Scene-Classification Dataset. It is a multiclass balanced dataset with the following 6 classes:

1. Buildings
2. Forests
3. Mountains
4. Glaciers
5. Streets
6. Sea

The dataset contains about ~17k labelled images from a wide range of natural scenes from all around the world. The task is to identify which kind of scene can the image be categorised into. The image dimensions are 150 x 150. The images are resized to 224x224, to be trained with AlexNet and VGG-16 for comparison.

Methodology

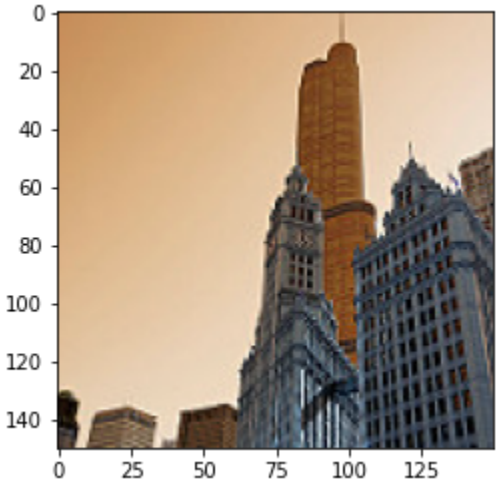
In this project, a comparative study of conventional CNNs and Interpretable CNNs is performed by training 2 versions of AlexNet and VGG-16 on the Scene-Classification dataset. The first version is the standard architecture known to all, while the second version modifies the architecture to include Convolution Mask layers, which will help to introduce Interpretability to this model.

We have also included helpful visualisations of higher level features to precisely showcase how the interpretable CNN learns important features compared to its conventional variant. The high level features are learnt irrespective of their position in the image.

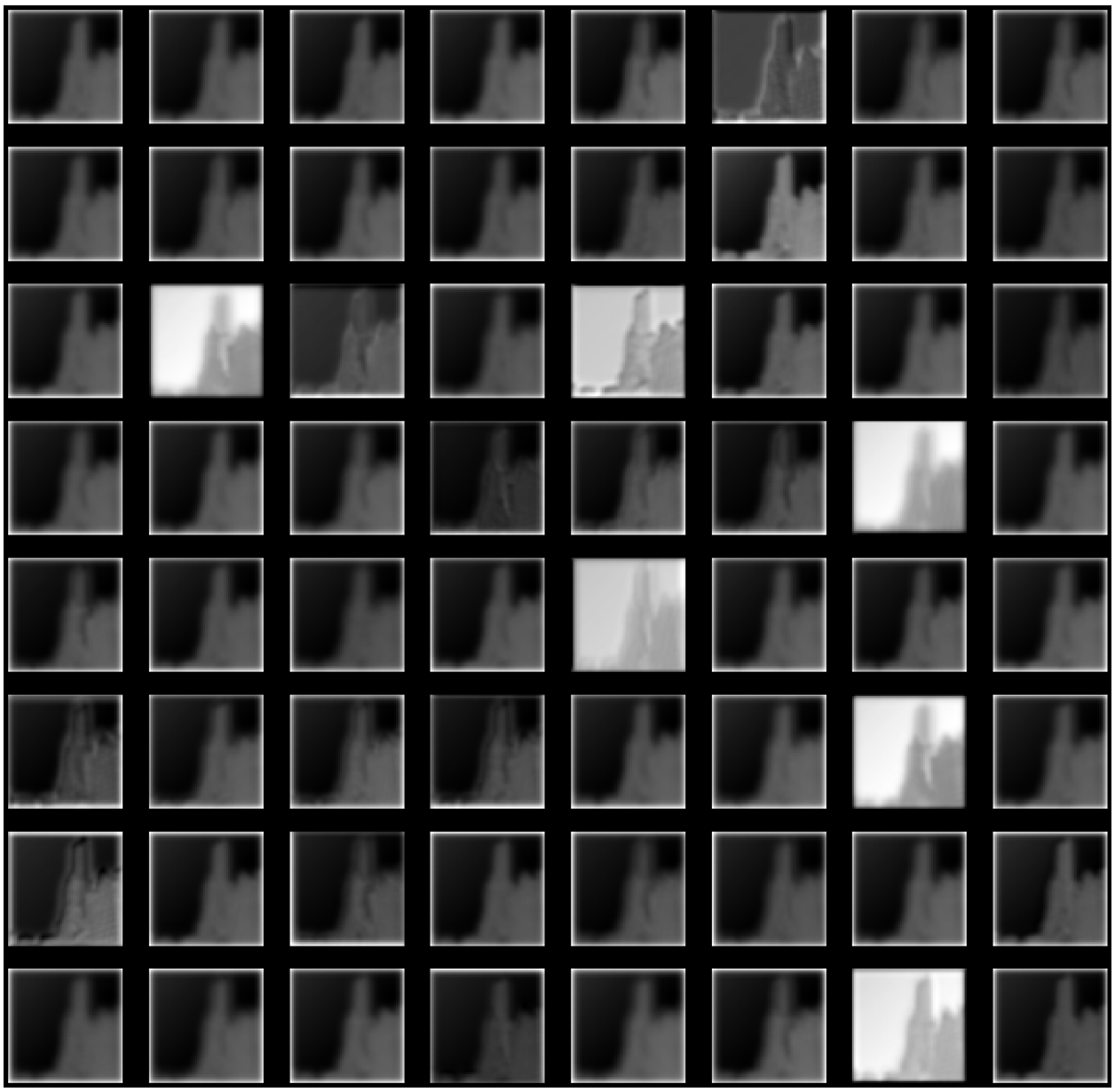
Visualisation

As can be seen in the following figures, the modified network architecture which includes Convolution_Mask layers learns distinct high level features, like the skyscraper present in the image, to do the classification

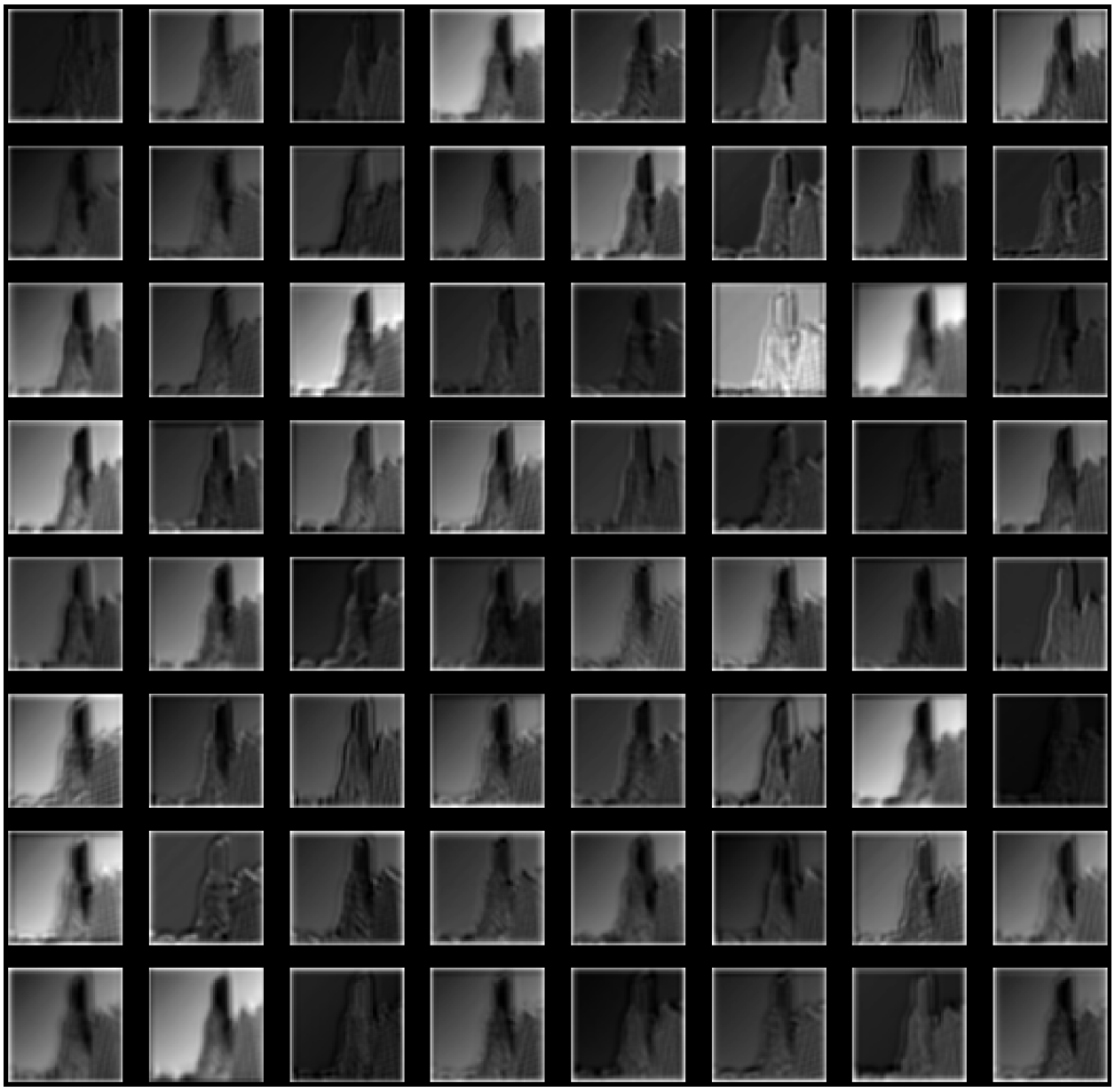
Input Image :



Visualisation of standard CNN:



Visualisation of Interpretable CNN:



Results of Comparative study

	Conventional	With Convolution Mask
AlexNet	90%	100%
VGG-16	91.7%	100%

Due to the low computational resources available to us and the large amounts of time required to train the model, we have used pre-trained weights to showcase a fair comparison. Due to pretrained weights and the small size of the dataset, the models achieve extremely high accuracies, possibly due to overfitting by the Convolution Mask layers. The presence of dropout layers did not prevent overfitting. However, the increase in accuracy for the interpretable models is inline with the results obtained in the paper. The higher accuracies show that the modified loss function and augmented back propagation increase interpretability in CNNs, which is also evident from the visualisations.

Salient Features

The most important feature of this approach to interpretability is the new loss function which has been introduced. This new function pushes the filters to learn specific high level features, which are human interpretable. Another feature are the templates created to account for the fact that the high level might not be present in a fixed position in the image. The forward and backwards propagation have also been modified to account for these changes.

Contributions

Through this project, we have showcased the strength of Interpretable CNNs on the balanced Scene-Classification Dataset. The Interpretable CNN has been extended to work on multiclass datasets. Additionally, we demonstrate enhanced interpretability through illustrative visualisations of higher level features directly via the kernels.