DATS 6303

Deep Learning

Dr. Amir Jafari

Final Project: Individual Report

by

Tyler Wallett

1. <u>Introduction</u>.

The project consisted of developing different explainable artificial intelligence (XAI) models for each of the deep learning models learned in class, which consisted of the following pages:

- XAI MLP
- XAI LSTM
- XAI CONVOLUTIONAL
    - GRADCAM
    - OCCLUSION SENSITIVITY
    - RISE
    - INTEGRATED GRADIENTS
    - LIME
- XAI OBJECT DETECTION
- XAI OFFERING

2. <u>Description of your individual work</u>.

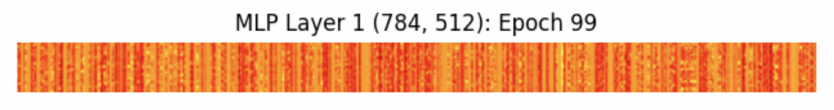
My work consisted of the parts highlighted in yellow above.

3. <u>Describe the portion of the work that you did on the project in detail</u>.

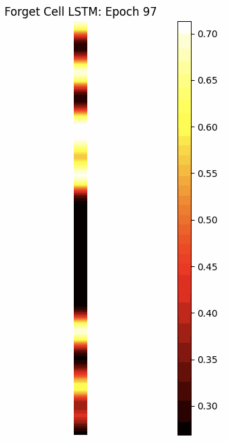
- XAI MLP consisted of a animated heatmap of the class activations in order to detect potential overfitting in the network.
- XAI LSTM consisted of an animation of the ‘forget gate’ in order to visualize the portion of the time sequence that the network is considering relevant to remember or forget.
- XAI CONVOLUTIONAL: GRADCAM consisted of using existing packages of XAI methods called Xplique, and creating a heatmap of the weights aggregated at the last feature map of an image classification model.
- XAI CONVOLUTIONAL: OCCLUSION SENSITIVITY consisted of using existing packages of XAI methods called Xplique, and creating different image representations of the image passed by the user by occluding several patches of the image, with a specific stride, in order to determine which patches of the image are the most relevant for the image classification.

4. <u>Results</u>.

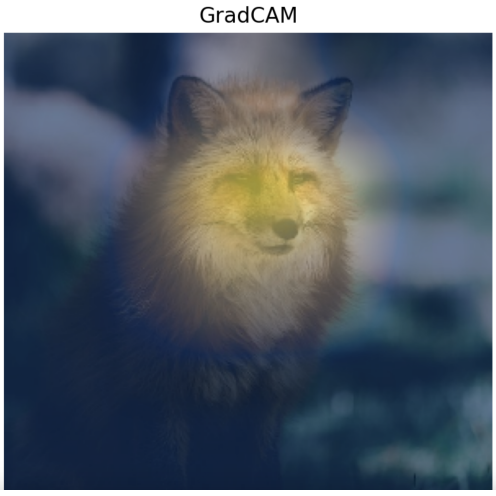
- XAI MLP: we can observe the straight opaque lines as zero vectors, which indicate a clear case of overfitting. Thus, in this manner, students can check and refer to whether their model is overfitting by looking at this animation.



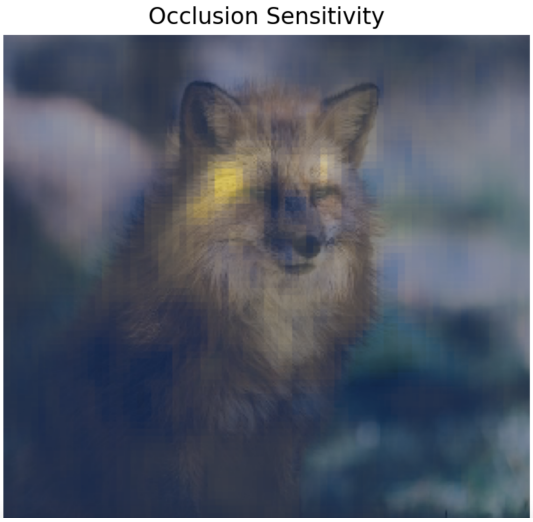
- XAI LSTM: we can see that the bright spots of the forget vector are remembering, as opposed to the opaque spots which are forgetting. In this manner we can clearly explain what is happening inside a LSTM network.



- XAI CONVOLUTIONAL GRADCAM: we can clearly observe, by the highlighted part in yellow, that the face of the fox is the most important region for the classification of the fox. In this manner we can clearly explain how the image classification model is concluding on its decision.



- XAI CONVOLUTIONAL OCCLUSION SENSITIVITY: we can clearly notice that the most important patches of the images are colored in yellow (Image size 299, Patch size 12x12, Patch stride 4x4) In this manner, we can identify the portions of the image that are relevant to the classification of the fox.



5. <u>Summary and conclusions</u>.

For the most part, the project was successful in delivering XAI models to users of the Streamlit application. However, the project's initial purpose was not successful as we attempted to devise fake images of faces using GANs in order to demonstrate whether we could explain that these images were fake using XAI models. A lot of time was spent in properly setting up the GANs, and the idea was abandoned. Thus, if we had more time perhaps we would have created our own solution without the need to refer to Xplique's existing solutions.
Had fun.

6. <u>Calculate the percentage of the code that you found or copied from the internet</u>.

50% was copied from the internet (GRADCAM and OCCLUSION SENSITIVITY)
50% was created by me using Pytorch and Matplotlib.pyplot (MLP and LSTM)

7. <u>References</u>.

For GradCAM:
https://deel-ai.github.io/xplique/latest/api/attributions/methods/grad_cam/
https://arxiv.org/pdf/1610.02391.pdf
For Occlusion Sensitivity:
https://deel-ai.github.io/xplique/latest/api/attributions/methods/occlusion/
https://arxiv.org/pdf/1711.06104.pdf