MULTI ATTENTION CONVOLUTIONAL NEURAL NETWORK

Jash Patel (M22CS061)   Bhawna Bhoria (M22MA003)

INTRODUCTION:

Fine-grained image recognition is able to tell the difference between things that only look slightly different, like different types of birds or car models. Traditional CNNs have trouble with this job because they focus on high level features and don't pay attention to details. The multi-attention network uses convolutional layers to put attention on different parts of a picture based on how important they are. This is done by learning attention maps, which show where important parts of a picture are. The network learns different attention maps by dividing feature maps into branches and using different attention methods on each branch. Each branch's attention map is put together to make a unified attention map. The network picks up small details, which makes it easier to tell the difference between things that look similar and makes it better than other CNN designs.

DATASET USED:

The CUB_200_2011 dataset is a benchmark dataset for fine-grained image recognition, consisting of approximately 11,000 images of 200 bird species. It is carefully annotated with bounding box annotations, part annotations, and attribute labels. The dataset is challenging due to the fine-grained nature of the task, with subtle visual differences between bird species. It includes a variety of bird species with different poses, appearances, and backgrounds. The training set has images, allowing models to learn discriminative features, while the testing set evaluates the generalization performance on unseen bird species.

Breakdown of the architecture:

SELayer:

- This class defines a module for the Squeeze-and-Excitation (SE) block.
- It takes as input the number of channels in the feature maps and a reduction ratio (defaulted to 1).
- The module performs adaptive average pooling and applies a series of fully connected layers with non-linear activations (Tanh and Sigmoid) to generate attention maps.
- The attention maps are used to weight the input feature maps, which are then returned.

MACNN:

- This class defines the overall MACNN model.
- It starts with a VGG19 backbone (pre-trained on ImageNet) to extract the feature maps from the input images.

- The SELayer is applied to four different feature maps extracted at different stages of the backbone network (se1, se2, se3, se4).
- Adaptive average pooling is used to obtain a global feature representation from the feature maps.
- Fully connected layers (fc1, fc2, fc3, fc4) are applied to the weighted feature maps from each SELayer to predict fine-grained class labels.
- The feature maps, global feature representation, attention maps, and predictions are concatenated and passed through another fully connected layer (fcall) to obtain the final prediction.

The forward method takes an input image and returns the following outputs:

- feat_maps: Feature maps extracted from the VGG19 backbone.
- cnn_pred: Prediction obtained by applying a fully connected layer to the global feature representation.
- A list of attention maps (P1, P2, P3, P4) generated by the SELayer for each level of the feature maps.
- A list of attention maps (M1, M2, M3, M4) obtained by applying normalization and squeezing dimensions on the attention maps.
- A list of attention maps (y1, y2, y3, y4) generated by the SELayer for visualization purposes.
- A list of predictions (pred1, pred2, pred3, pred4, pred) obtained by applying fully connected layers to the weighted feature maps and the global feature representation.

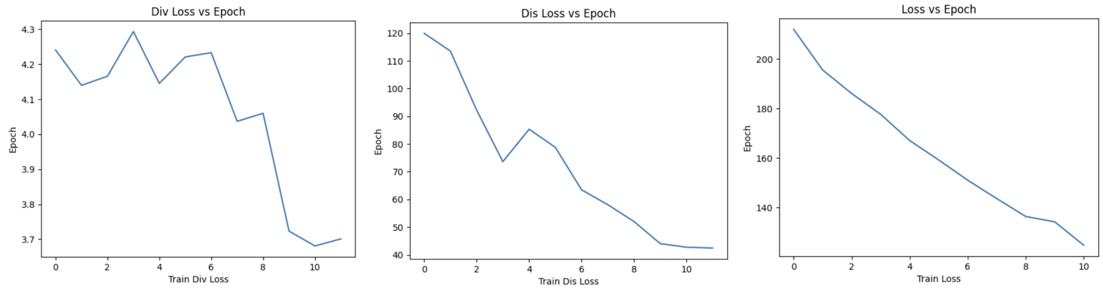
Overall, the MACNN model integrates attention mechanisms through SE blocks to capture fine-grained details and improve classification performance for the CUB_200_2011 dataset.

Model Architecture			
Operations	Input shape	Output shape	Parameters
VGG19	(3,224,224)	(512,14,14)	2,359,808(output)
AdaptiveAvgPool2d	(512,14,14)	(512,7,7)	-
SE1	(512,7,7)	(512,1,1)	262,144
SE2	(512,7,7)	(512,1,1)	262,144
SE3	(512,7,7)	(512,1,1)	262,144
SE4	(512,7,7)	(512,1,1)	262,144
Classifier Layer 1	(16,16,512)	(200)	512,200
Classifier Layer 2	(16,16,512)	(200)	512,200
Classifier Layer 3	(16,16,512)	(200)	512,200
Classifier Layer 4	(16,16,512)	(200)	512,200
AdaptiveAvgPool2d	(512,1,1)	(512,1,1)	-
Final Classifier Layer	(16,16,512)	(200)	512,200

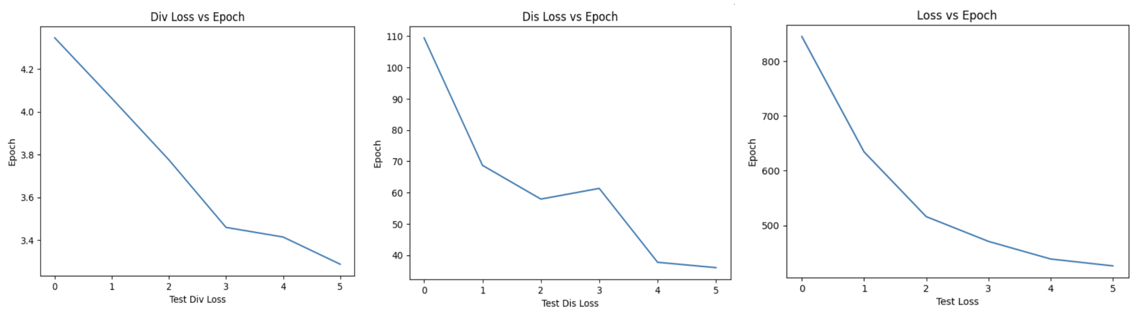
RESULTS:

The following results are observed when the model is trained for 30 epochs with SGD optimizer with learning rate of 10^{-3}. The performance of the model is obtained by accuracy and loss functions. Div Loss, Dis Loss and Cross Entropy Loss.

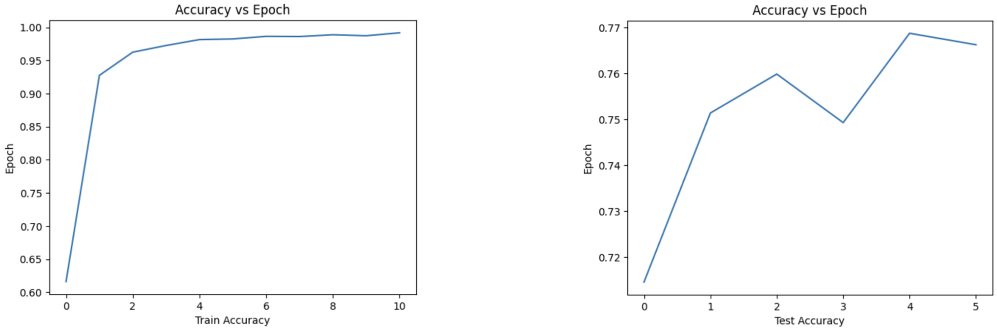
Train Loss Curves:-



Test Loss Curves:-

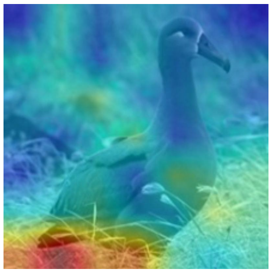


Train vs Test Accuracy:-

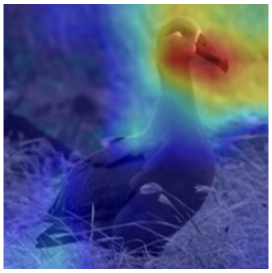


Original Image

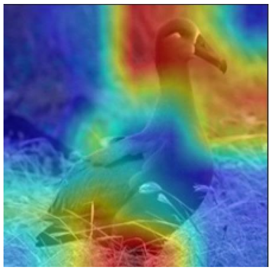
Untrained w/o Attention



Trained w/o Attention



Trained with Attention

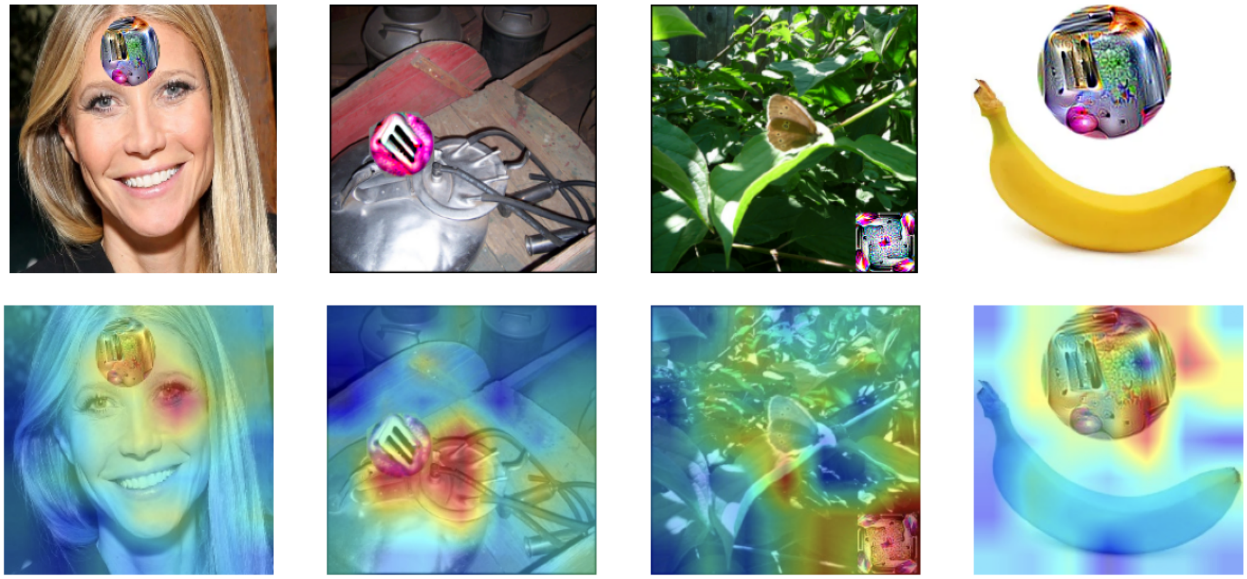


The key DL parts of the MACNN can be summarized as follows:

Multi-Attention Convolutional Neural Network (MACNN) designed to classify bird species in the CUB_200_2011 dataset. The MACNN consists of a modified VGG-19 architecture and four Squeeze-and-Excitation (SE) modules.
The SE modules, implemented in the SELayer class, are inserted between the convolutional layers of the VGG-19 architecture to adaptively recalibrate the feature maps. Each SE module takes as input the feature maps from a specific convolutional block of the VGG-19 architecture and produces an attention map, a modulation map, and a feature descriptor. The attention map highlights the important regions of the feature maps, the modulation map weights the contribution of the feature maps to the classification task, and the feature descriptor is a learned representation of the feature maps.
The output of the SE modules is then concatenated with the global average pooling of the feature maps and fed into a fully connected layer to make the final prediction. The MACNN also includes four separate fully connected layers that receive as input the feature descriptors from each SE module. This allows the MACNN to attend to multiple levels of abstraction in the feature maps and make more informed predictions.
In summary, the MACNN model architecture consists of a modified VGG-19 backbone, four SE modules, and five fully connected layers. The SE modules provide attention and modulation maps to adaptively recalibrate the feature maps, while the fully connected layers combine the feature descriptors from the SE modules to make the final prediction.

Observation:-

Multi Attention Model focuses on adversarial patches.



References:

https://towardsdatascience.com/bird-by-bird-using-deep-learning-4c0fa81365d7
https://openaccess.thecvf.com/content_ICCV_2017/papers/Zheng_Learning_Multi-Attention_Convolutional_ICCV_2017_paper.pdf
https://www.vision.caltech.edu/datasets/cub_200_2011/