Design Credits Project Report

Naikawadi Shivaprasad (B20AI057)

Kethireddy Harshith Reddy (B20AI018)

Problem Statement

Classifying objects in drone imagery is hard because of how different the objects look and how complicated the background is. Traditional classification methods frequently use hand-made characteristics without taking into account the objects' surroundings. So, an attention-based approach to classifying objects was made, which uses deep learning models to automatically learn distinguishing features and pay attention to relevant parts of an image. The goal is to make it easier and more accurate to classify objects in images taken by drones for uses like surveillance, mapping, and inspection.

Motivation

The motivation is to use attention mechanisms that can help improve the accuracy and efficiency of object classification by focusing on the most relevant parts of the image rather than processing the entire image. In the context of drone imagery, attention mechanisms can be particularly useful for identifying objects that are partially obscured or in difficult-to-see locations, such as cars parked under trees or street lamps in the shadows. This improved classification can have numerous practical applications, such as improving traffic monitoring and management, enhancing urban planning and development, and aiding in disaster response and recovery efforts.

Proposed Methodology

The attention mechanism is a proposed method that adds back features that were lost and uses convolution layers to pull out more diverse features. Convolutional neural networks were used to find or classify things on the ground. They were compared with some existing methods and the proposed attention blocks. The proposed attention-based CNN architecture was adopted and compared comprehensively with the existing networks like *MobileNet, VGG16*, and *ResNet*.

Dataset and its characteristics

The VisDrone2019 dataset, which was put together by the AISKYEYE team at the Lab of Machine Learning and Data Mining, was used. This dataset includes:
- 6471 training samples
- 548 validation samples
- 1610 test samples

Some sample images from the dataset:

There are many things about the drone dataset. In the datasets we already have, objects are taken by people or by CCTV. A circle around a car is a path that can be taken based on the car. So, you can mostly see the front, back, and sides of the object, and you can also see a little bit of the top. Because the picture was taken up close, it is big and clear.

On the other hand, drones can take pictures in a variety of car-shaped, semi-spherical shapes. It would have more information than traditional datasets, especially from a bird's-eye view and a plane. So, the object looks the same from the front, back, side, and top. But it looks different when taken at a 90-degree angle. Because they can only see the top, people look like dots, and street lamps look like straight lines. In this case, it's hard to put things into groups, which makes the classification task difficult.
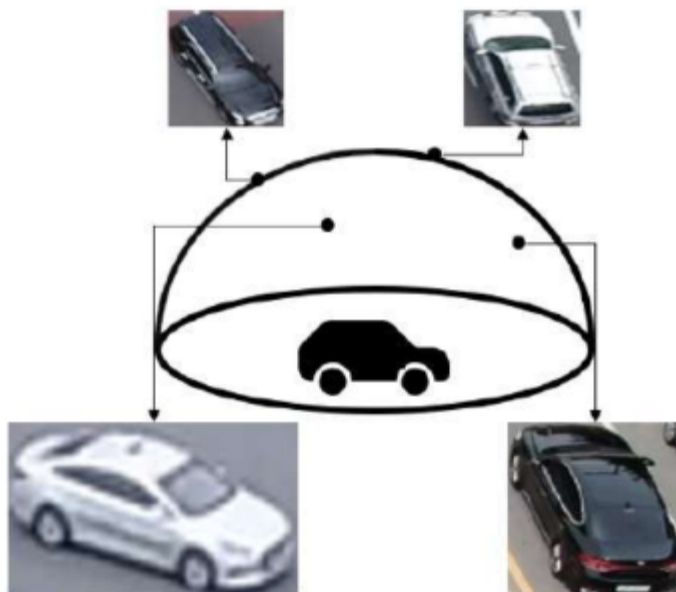
In the picture above, we can see that the shape of the cars changes depending on how the picture was taken. Because there are images from different angles and at different resolutions, it is hard to put them into groups.

Attention Block

Attention block is a proposed technique that complements lost features and extracts more diverse features through convolution layers. A feature map called attention block is added to the output of the convolution layer before entering the next convolution layer's input.
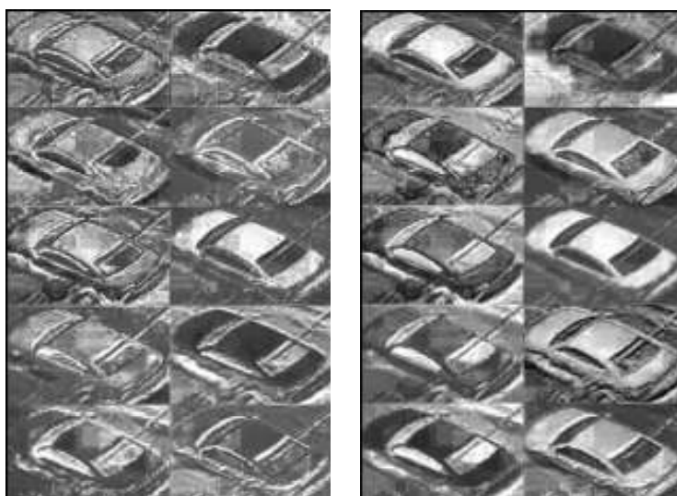
In the image above, the class on the image is a car, and the initial width and height are about 100 pixels each. The picture on the left shows what happened when it was shrunk to 224x224 and put into the convolution layer. When you add the attention block to the picture on the left, you get the image on the right. In the picture on the right, you can clearly see what happens when the attention block is used. So, as the network gets deeper, continuing to feed it features from the original picture makes the learning process better.
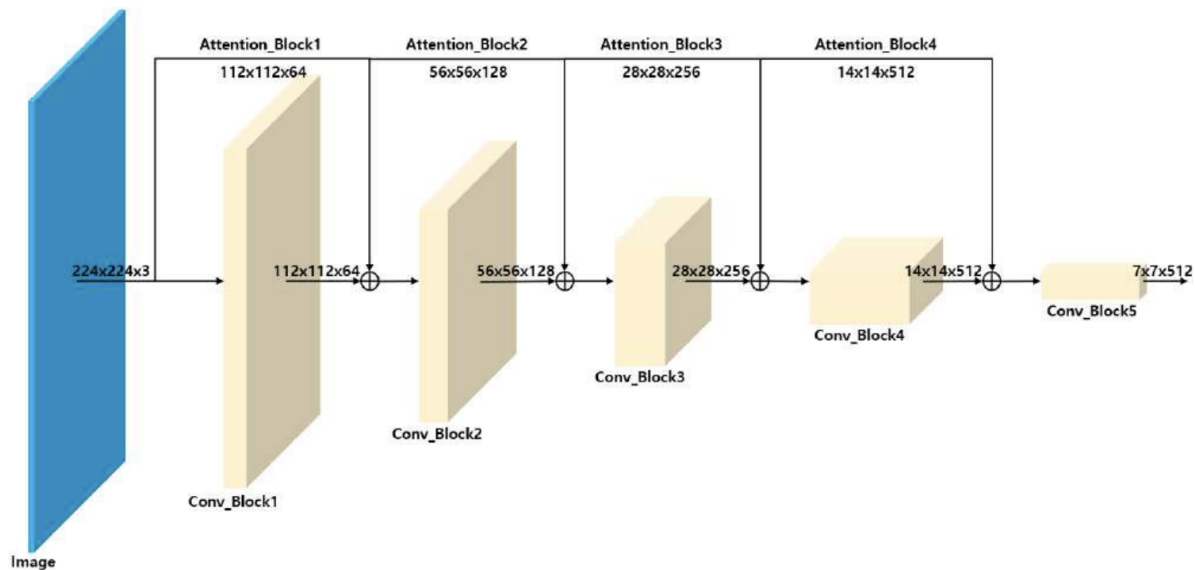
In this project, convolutional and attentional blocks are used to make networks. The internal structure of the convolution block is simple because it is made up of only four convolution layers and one max-pooling layer. In the convolution layer part, the bottleneck method is used on two layers with kernel sizes of 33 and two layers with kernel sizes of 11. This method is used because the bottleneck technique keeps the same speed while reducing the number of parameters by about 28%. The output of the convolution block will be a quarter the size of the picture that went into it, and the number of channels will double.

The attention block has a simple structure on the inside as well. At first, it had just 11 convolution layers and one max-pooling layer. Each layer of the attention block goes up by one layer as the convolution block goes deeper. It only helps to make a new feature map by making the number of channels bigger and the size of the picture smaller. The reason for setting up as stated is to keep as many of the properties of the input picture as possible and to reduce the amount of work that needs to be done. After these outputs are added, they are sent to the next convolution layer and continue to be learned through batch normalization and activation functions.

Layer	Output
Input Image	224 x 224 x 3
Conv_Block1	112 x 112 x 64
Attention_Block1	112 x 112 x 64
Conv_Block1 + Attention_Block1	112 x 112 x 64
Conv_Block2	56 x 56 x 128
Attention_Block2	56 x 56 x 128
Conv_Block2 + Attention_Block2	56 x 56 x 128
Conv_Block3	28 x 28 x 256
Attention_Block3	28 x 28 x 256
Conv_Block3 + Attention_Block3	28 x 28 x 256
Conv_Block4	14 x 14 x 512
Attention_Block4	14 x 14 x 512
Conv_Block4 + Attention_Block4	14 x 14 x 512
Conv_Block5	7 x 7 x 512

The structure of the image showing the architecture of the model is described in the table above.

Performance comparison of models

Model	Test Accuracy
CNN with Attention	93.87%
MobileNet	89.62%
VGG16	82.87%
Resnet	88.50%
CNN	81.37%

Performance curves
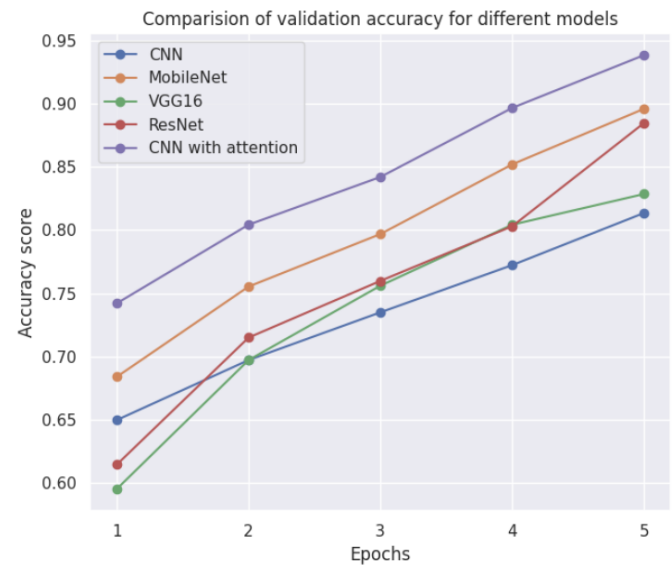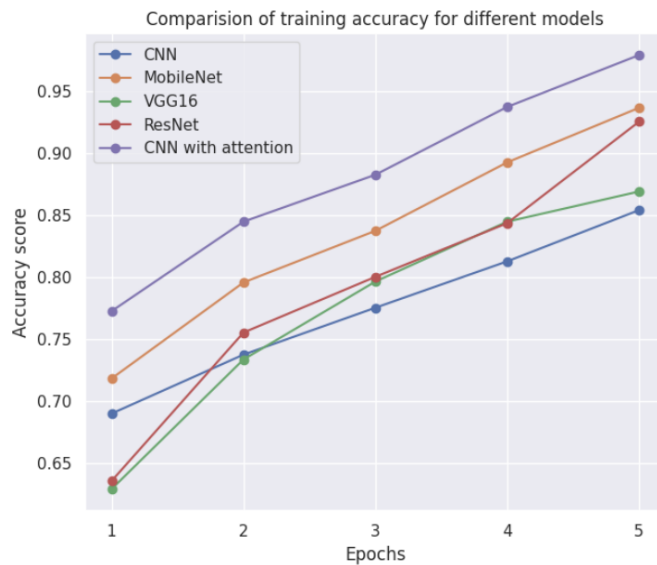
Comparison of parameters and model size

Model	Parameters	Model Size
CNN with Attention	11,177,538	45.43 MB
MobileNet	8,251,397	42.71 MB
VGG16	134,268,738	512.23 MB
Resnet	25,557,032	95.45 MB
CNN	23,653,314	90.33 MB

From the results above, we can see that the CNN model's accuracy has improved with the attention mechanism. When we compared the trainable parameters and the model size of each model, we saw that CNN with Attention has fewer trainable parameters and a smaller model size than other models, except for MobileNet, which has fewer values than CNN with Attention but less accuracy than CNN with Attention (proposed), making it the better model.

Conclusion

- Attention mechanisms have been shown to be effective in improving the performance of deep neural networks for various computer vision tasks, including object classification, by allowing the network to focus on the most relevant parts of the input.
- We've implemented an Attention mechanism on the VisDrone19 dataset.
- We also compared the results of evaluating various models on this dataset to analyze the contribution of different components in our attention mechanism and provide insights into the underlying mechanism of attention in object classification.
- The experimental results demonstrated that our attention mechanism achieved significant improvements in accuracy and robustness and outperformed the baseline methods.
- Our proposed attention mechanism has the potential to be applied to other computer vision tasks for drone imagery and to facilitate the development of more advanced and intelligent drone systems.

Code

- The code is present on the server at the path */users/DC_Project*. You can execute the code in any text editor (like VS Code) by installing the extension for .ipynb as defined in the README file.
- Otherwise, you can just follow the README file to execute the code locally on your PC.

References

https://ieeexplore.ieee.org/document/9589099

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9589099