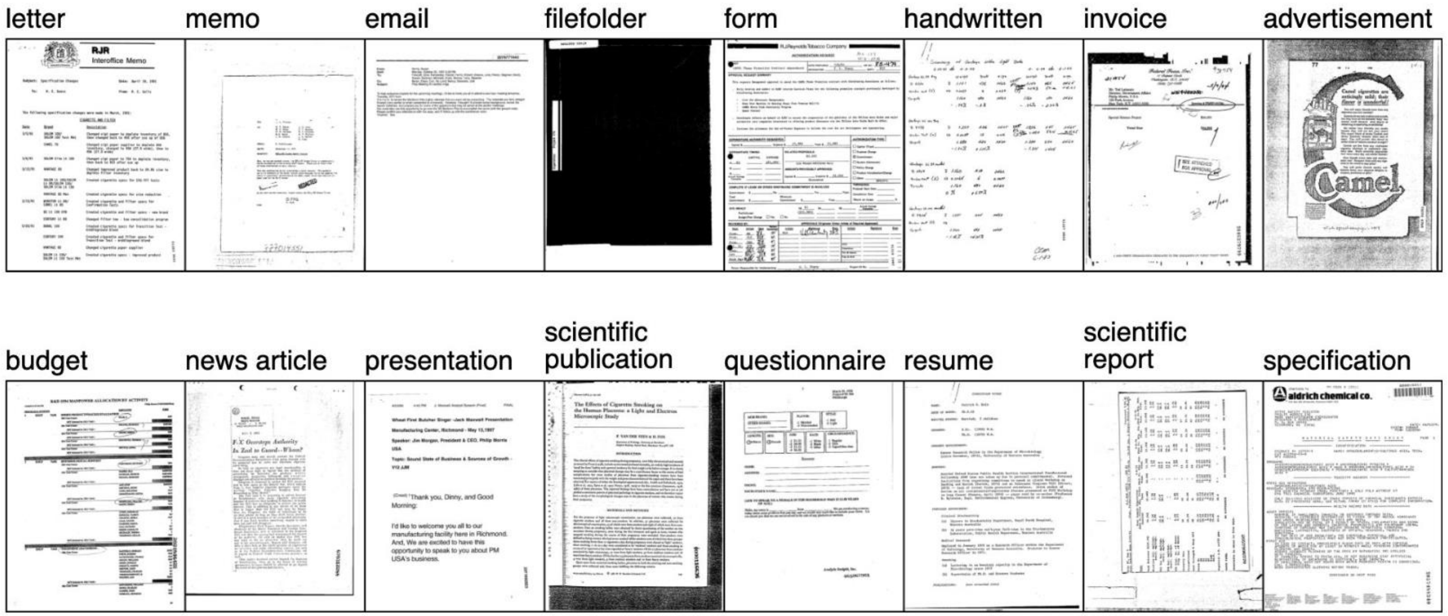
DOCUMENT CLASSIFICATION USING VGG-16 WITH ATTENTION

Submitted By Rishab Tomar (224161007) Atul Bhagat(224151002)

Submitted To Dr. Debanga Raj Neog

ABOUT DATASET

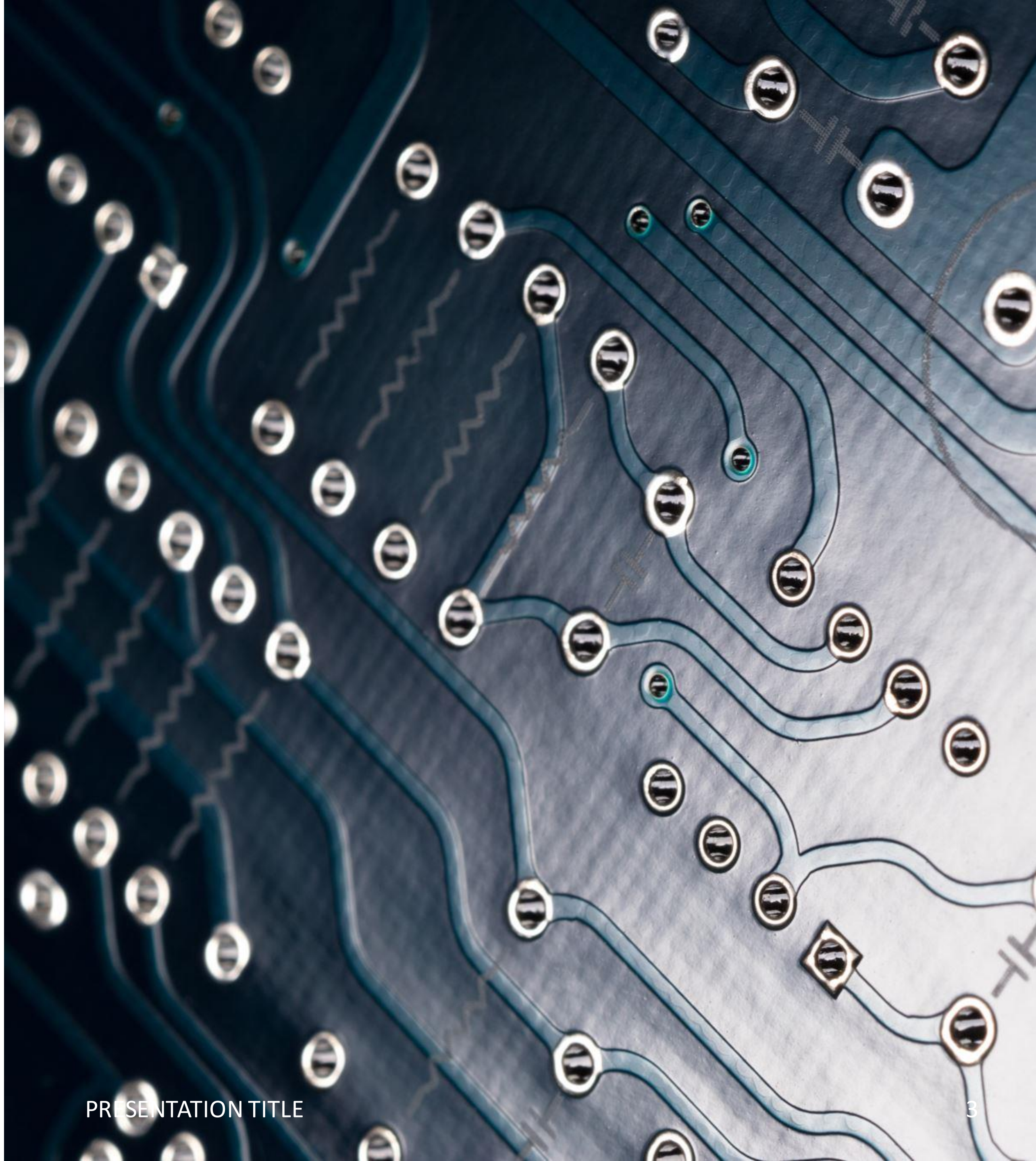


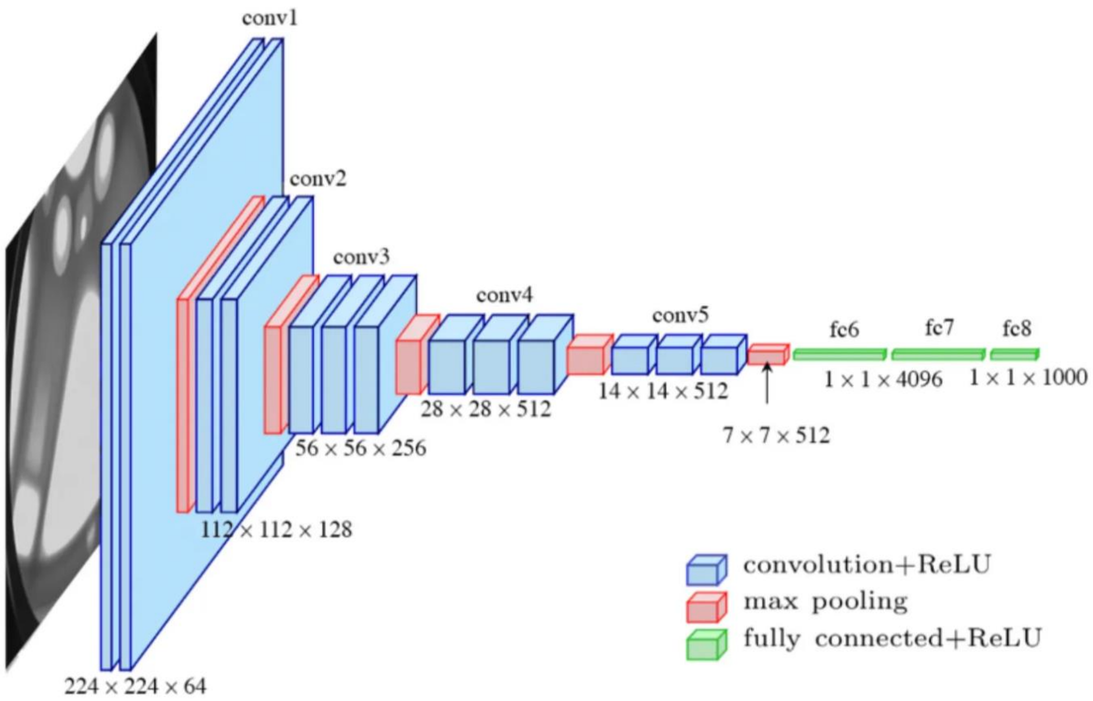
- RVL-CDIP Dataset containing a total of 16 document categories including scientific articles ,handwritten letters. newspapers and more.

- It Contains a total of 400000 images 320000 as Training and 40000 as Validation and 40000 as Testing, but we have used only 16000 images with 1000 images per class to maintain a balance class distribution among classes.

PROBLEM STATEMENT

- We aim to explore and evaluate the impact of incorporating an attention block into a VGG-16 network compared to a VGG-16 network without attention.
- The RVL-CDIP dataset used contains a diverse collection of images, and the task is to classify them into specific categories

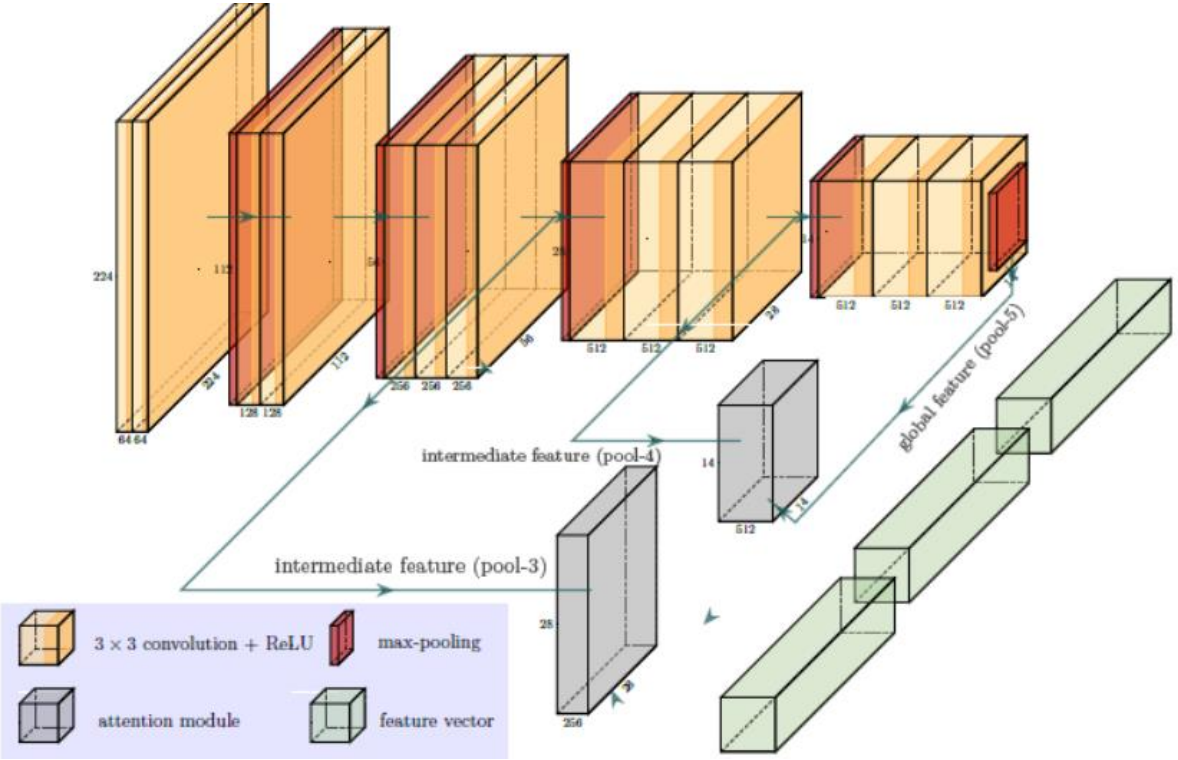




MODEL-VGG16

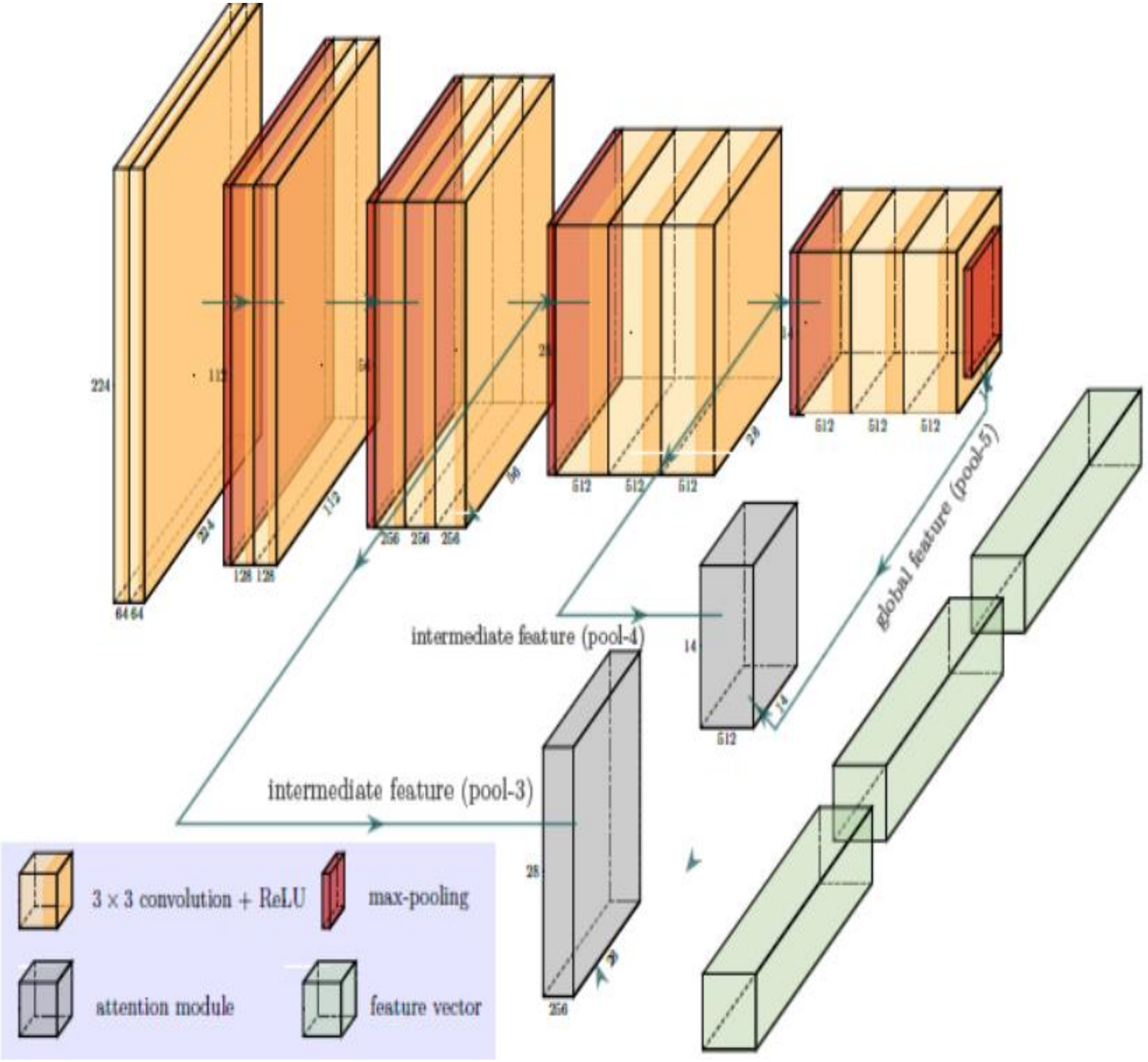
- VGG-16, short for the Visual Geometry Group 16-layer network, is a deep convolutional neural network architecture designed for image classification and object recognition tasks.
- VGG-16 consists of 16 layers, including 13 convolutional layers and 3 fully connected layers. It is characterized by its use of small 3x3 convolutional filters and max-pooling layers, which result in a highly expressive and deep network

MODEL-VGG16 WITH ATTENTION



- Two attention modules are applied (the gray blocks). The output of intermediate feature maps(pool-3 and pool-4) are used to infer attention maps. Output of pool-5 serves as a form of global-guidance because the last stage feature contains the most abstract and compressed information over the entire image.

- The three feature vectors (green blocks) are computed via global average pooling and are concatenated together to form the final feature vector, which serves as the input to the classification layer



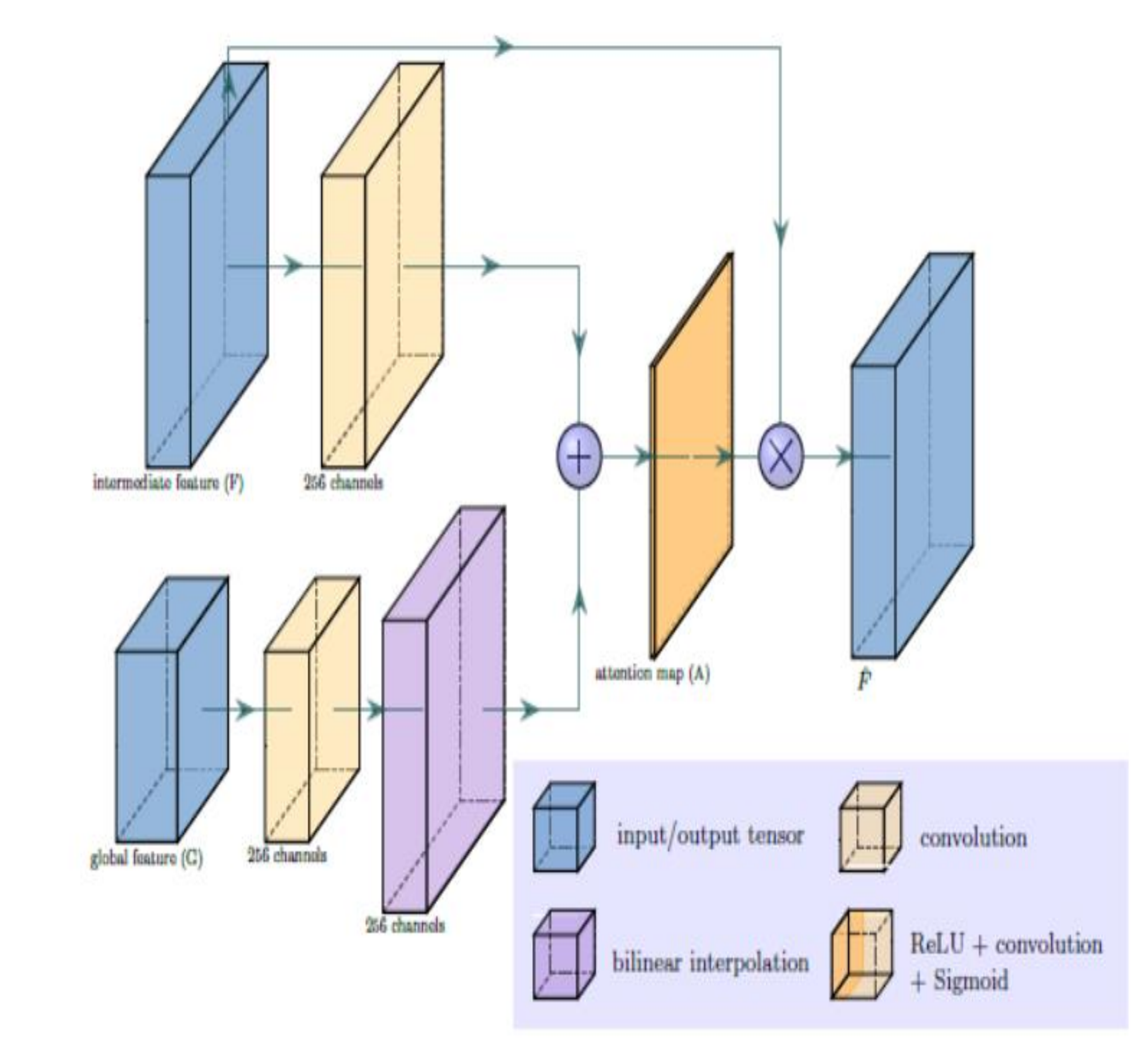
- The intermediate feature vector(F) is the output of pool-3 or pool-4 and the global feature vector (output of pool-5) is fed as input to the attention layer.

- Both the feature vectors pass through a convolution layer.

- If spatial size of global and intermediate features are different, feature upsampling is done via bilinear interpolation.

- The *up_factor* determines by what factor the convoluted global feature vector has to be upscaled.

- After that an element wise sum is done followed by a convolution operation that just reduces the 256 channels to 1.



- This is then fed into a Softmax layer, which gives us a normalized Attention map (A). Each scalar element in A represents the degree of attention to the corresponding spatial feature vector in F.

- The new feature vector $\hat{F}$ is then computed by *pixel-wise* multiplication. That is, each feature vector f is multiplied by the attention element a

- So, the attention map A and the new feature vector $\hat{F}$ are the outputs of the Attention Layer.

- The architecture of VGG16 is kept mostly the same except the Dense layers are removed.

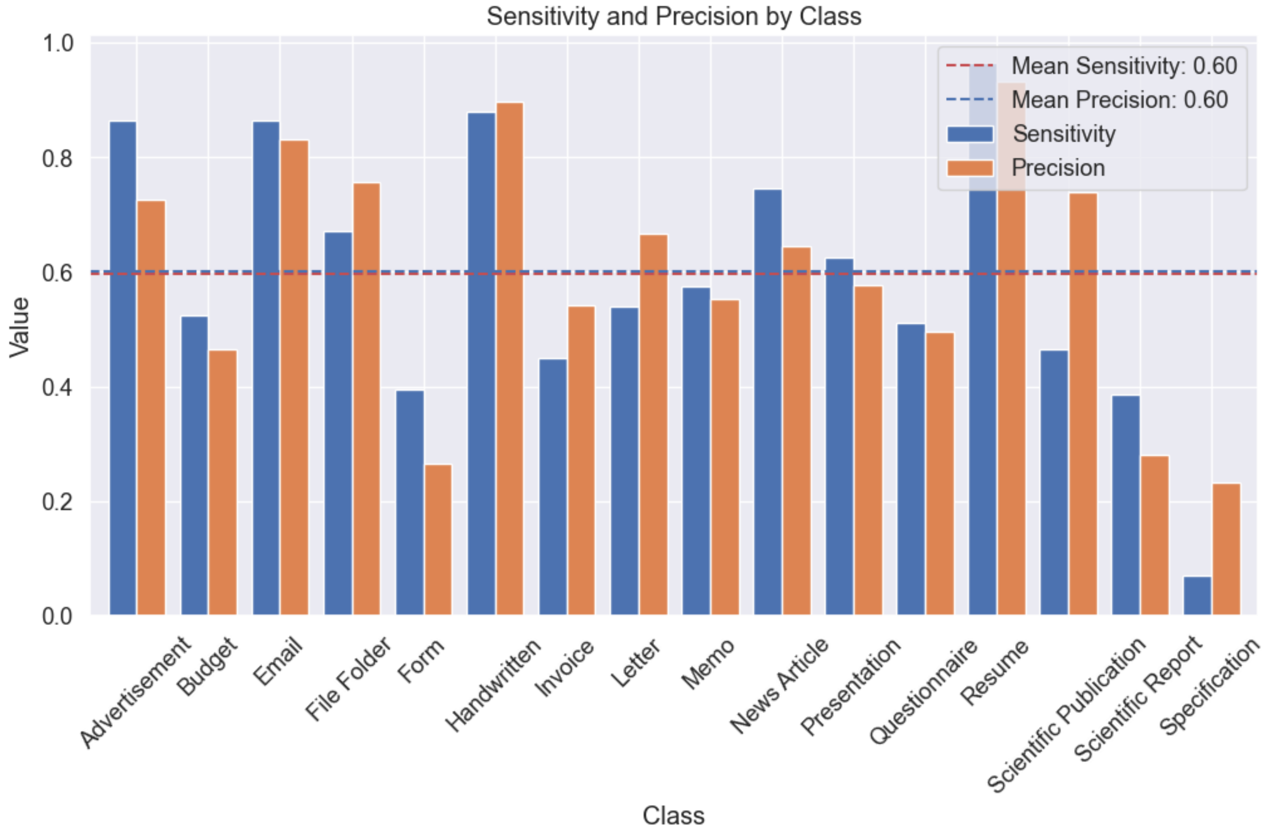
- We pass pool-3 and pool-4 through the attention layer to get $\hat{F}3$ and $\hat{F}4$.

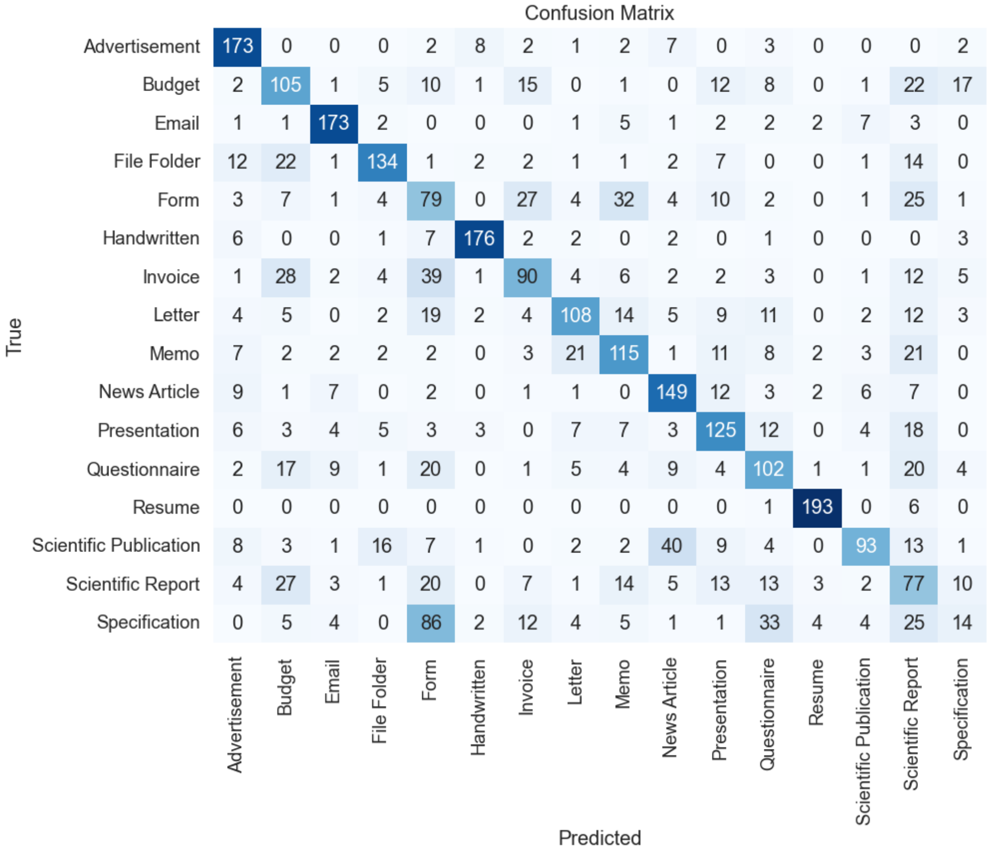
- $\hat{F}3$,$\hat{F}4$ and G(pool-5) are concatenated and fed into the final classification layer.

- The whole network is trained end-to-end.

RESULTS

1. VGG-16

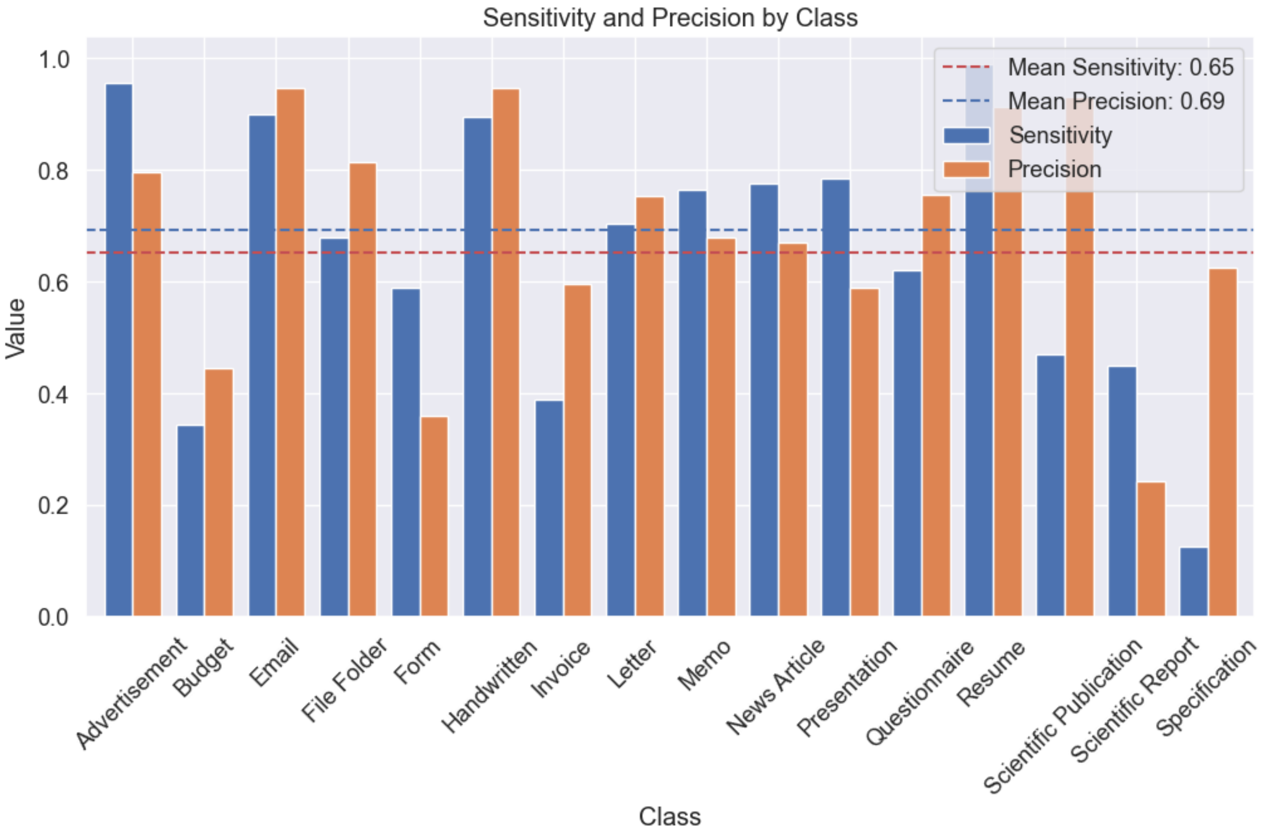


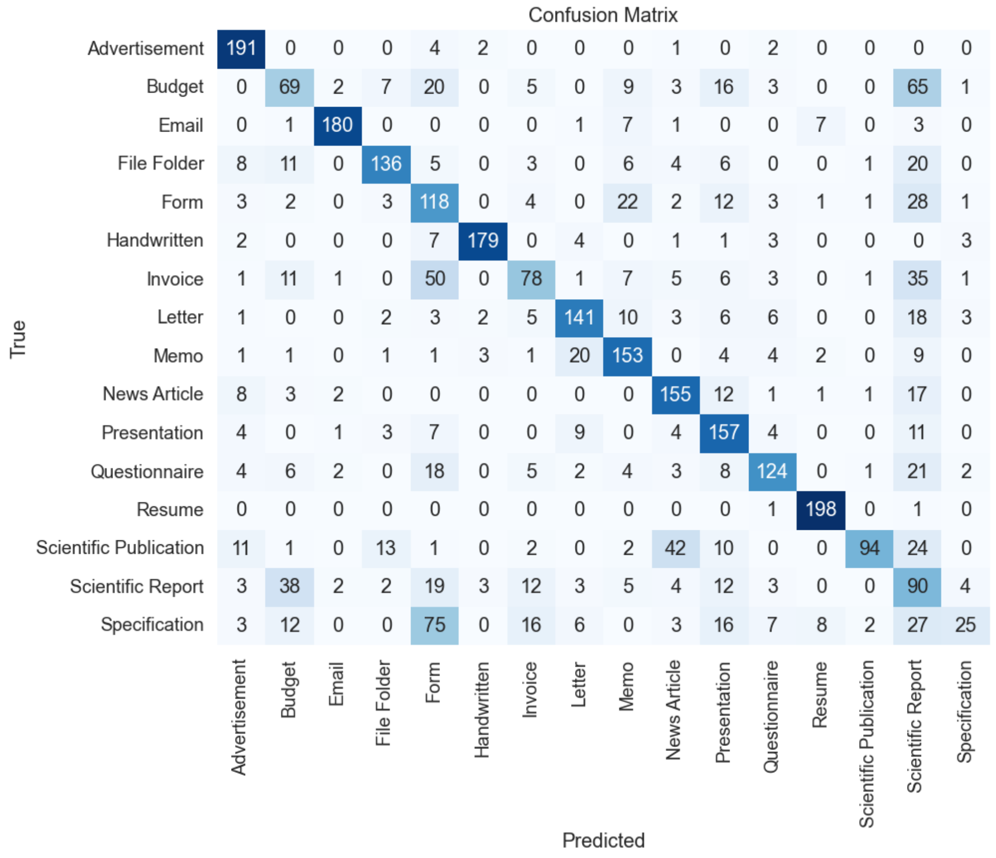


Confusion Matrix

	Advertisement	Budget	Email	File Folder	Form	Handwritten	Invoice	Letter	Memo	News Article	Presentation	Questionnaire	Resume	Scientific Publication	Scientific Report	Specification
Advertisement	173	0	0	0	2	8	2	1	2	7	0	3	0	0	0	2
Budget	2	105	1	5	10	1	15	0	1	0	12	8	0	1	22	17
Email	1	1	173	2	0	0	0	1	5	1	2	2	2	7	3	0
File Folder	12	22	1	134	1	2	2	1	1	2	7	0	0	1	14	0
Form	3	7	1	4	79	0	27	4	32	4	10	2	0	1	25	1
Handwritten	6	0	0	1	7	176	2	2	0	2	0	1	0	0	0	3
Invoice	1	28	2	4	39	1	90	4	6	2	2	3	0	1	12	5
Letter	4	5	0	2	19	2	4	108	14	5	9	11	0	2	12	3
Memo	7	2	2	2	2	0	3	21	115	1	11	8	2	3	21	0
News Article	9	1	7	0	2	0	1	1	0	149	12	3	2	6	7	0
Presentation	6	3	4	5	3	3	0	7	7	3	125	12	0	4	18	0
Questionnaire	2	17	9	1	20	0	1	5	4	9	4	102	1	1	20	4
Resume	0	0	0	0	0	0	0	0	0	0	0	1	193	0	6	0
Scientific Publication	8	3	1	16	7	1	0	2	2	40	9	4	0	93	13	1
Scientific Report	4	27	3	1	20	0	7	1	14	5	13	13	3	2	77	10
Specification	0	5	4	0	86	2	12	4	5	1	1	33	4	4	25	14

2. VGG16 WITH ATTENTION





	Advertisement	Budget	Email	File Folder	Form	Handwritten	Invoice	Letter	Memo	News Article	Presentation	Questionnaire	Resume	Scientific Publication	Scientific Report	Specification
Advertisement	191	0	0	0	4	2	0	0	0	1	0	2	0	0	0	0
Budget	0	69	2	7	20	0	5	0	9	3	16	3	0	0	65	1
Email	0	1	180	0	0	0	0	1	7	1	0	0	7	0	3	0
File Folder	8	11	0	136	5	0	3	0	6	4	6	0	0	1	20	0
Form	3	2	0	3	118	0	4	0	22	2	12	3	1	1	28	1
Handwritten	2	0	0	0	7	179	0	4	0	1	1	3	0	0	0	3
Invoice	1	11	1	0	50	0	78	1	7	5	6	3	0	1	35	1
Letter	1	0	0	2	3	2	5	141	10	3	6	6	0	0	18	3
Memo	1	1	0	1	1	3	1	20	153	0	4	4	2	0	9	0
News Article	8	3	2	0	0	0	0	0	0	155	12	1	1	1	17	0
Presentation	4	0	1	3	7	0	0	9	0	4	157	4	0	0	11	0
Questionnaire	4	6	2	0	18	0	5	2	4	3	8	124	0	1	21	2
Resume	0	0	0	0	0	0	0	0	0	0	0	1	198	0	1	0
Scientific Publication	11	1	0	13	1	0	2	0	2	42	10	0	0	94	24	0
Scientific Report	3	38	2	2	19	3	12	3	5	4	12	3	0	0	90	4
Specification	3	12	0	0	75	0	16	6	0	3	16	7	8	2	27	25

RESULTS DISCUSSION

- VGG16(Model-1) shows relatively good sensitivity and precision for most classes, with an average sensitivity of 0.5956 and an average precision of 0.6004.
- VGG16 With Attention(Model-2), on the other hand, generally has higher sensitivity and precision values across the classes, resulting in a slightly better performance with an average sensitivity of 0.6525 and an average precision of 0.6918.
- The Percentage Increase in the mean sensitivity is 9.56%, and the percentage increase in the mean precision is 15.24% from Model-1 to Model-2.
- The Attention layer in Model 2, used in conjunction with VGG16, enhances performance by selectively emphasizing important image regions, improving precision, reducing overfitting, and providing interpretability to the model's predictions. It optimizes feature extraction and object recognition