

Question1: Analyzing bias for different ML Algorithms

Dataset Used: UTK Face

1. Make images as gray scaled because LCNN prefers 1 channel image
2. Also resize the dataset images to 64 X 64
3. Age Demographic Subgroups (A1: [0,28], A2: [29,56], A3: [57,84], A4: [85,116])
4. Race Demographic Subgroups (R1, R2, R3, R4)
5. Gender Demographic Subgroups (G1: Male(0) , G2: Female(1))

Dataset Split:

Split the dataset into:

1. **Train set:** 70%
2. **Validation set:** 10%
3. **Test set:** 20%

Algorithm:

I. LCNN model (9 Convolution layer version)

A. Test Accuracy: 84.96 %

B. Confusion Matrix:

Predicted	Actual	
	1	0
1	1876	342
0	371	2152

II. Linear SVM

A. Test Accuracy: 83.21 %

B. Confusion Matrix:

		Actual	
		1	0
Predicted	1	1836	398
	0	411	2096

1. Testing Performance:

- Degree of Bias (DoB) = $\text{std}(CAcc_{D_j}), \forall j$
- Precise Subgroup Equivalence (PSE) = $\frac{(1-DI) + AFR + DoB}{3}$

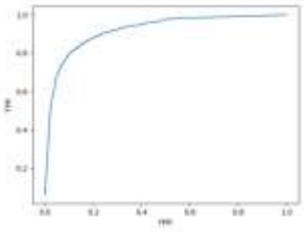
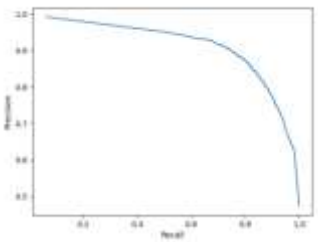
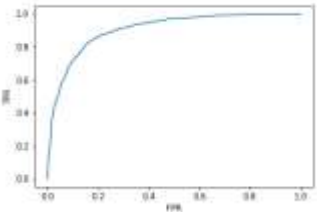
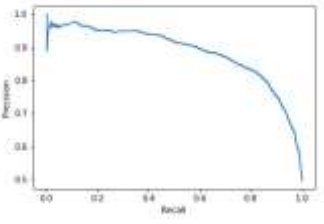
Table 1.1. Gender prediction across all Race categories (R1, R2, R3, R4)

Model	Classification Accuracy (in %)								DoB (in %)
	R1		R2		R3		R4		
	G1	G2	G1	G2	G1	G2	G1	G2	

Table 1.2. Gender prediction across all Age groups (A1, A2, A3, A4)

Model	Classification Accuracy (in %)								DoB (in %)
	A1		A2		A3		A4		
	G1	G2	G1	G2	G1	G2	G1	G2	
									</

2. & 3. ROC Curve & Precision-Recall Curve:

Model	ROC curve	Precision-Recall Curve
LCNN9		
Linear SVM		

4. Precision-Recall curve is more reliable for biased/imbalanced data, as both Precision and Recall focuses on the minority class and is unconcerned about the majority class.

As we can see Precision-Recall curve is focused towards correct prediction of positive class (minority class), makes it effective diagnostic for imbalanced binary classification model. Whereas ROC analysis doesn't have any bias towards the models that perform well on the minority class at the expense of the majority class.

- $Precision = TP / (TP + FP)$: It refers to positive predictive value. It tells how good the model is at predicting the positive class.
- $Recall = TP / (TP + FN)$: It refers to the number of correct positive predictive made out of all positive predictions that could have been made.

Analysis: We can observe from curves of Precision-Recall of LCNN9 and Linear SVM. Linear SVM is not able to classify properly as it leads to more number of False Positive predicted samples due to linear model and data distribution can be non-linear whereas LCNN has less number of False Positive predicted as it has activation functions used which leads to non-linearity compare to False Positives by Linear SVM. LCNN9 is a better model as compared to Linear SVM if we see it in terms of overall performance of model as LCNN9 model has higher accuracy than Linear SVM model. However when we go through the **Degree of Bias (DoB)**, LCNN9 has average DoB across Race and Age ($[2.708+9.124]/2 = 11.832$) whereas Linear SVM has average DoB across Race and Age ($[4.374+7.043]/2 = 11.417$). So on the basis of DoB (one of the Bias metric) we can say that Linear SVM is better as compared to LCNN9. But we can't conclude best model either on the basis of overall performance or DoB. We have to use **PSE(Precise Subgroup Equivalence)** metric which takes combination of **both overall performance as well as impact of Bias** to select the best model out of Linear SVM and LCNN9. So for that we need to calculate **Disparate Impact(DI)** and **Average False Rate(AFR)** also.

5. A. As we can visualize from *Table 1.1*. Gender prediction across various Race categories (R1,R2,R3,R4) LCNN9 model leads to less Bias as compared to Linear SVM model as LCNN9 has low Degree of Bias (DoB) as compared to Linear SVM.

B. As we can see from *Table 1.2*. Gender prediction across various Age groups (A1, A2, A3, A4) Linear SVM model leads to less Bias as compared to LCNN9 model as Linear SVM has low Degree of Bias (DoB) as compared to LCNN9.

6. Focal Loss

A. Test Accuracy: 85.382 %

B. Confusion Matrix:

		Actual	
		1	0
Predicted	1	1911	357
	0	332	2137

Table 6.1. Gender prediction across all Race categories (R1, R2, R3, R4)

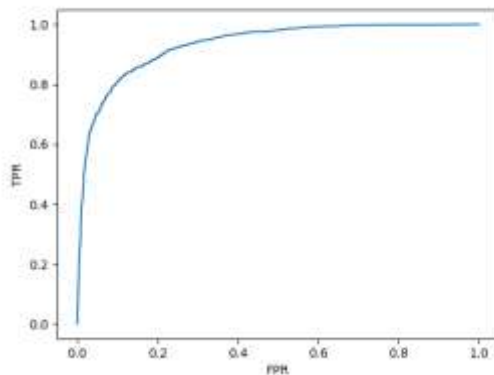
Model	Classification Accuracy (in %)								DoB (in %)
	R1		R2		R3		R4		
	G1	G2	G1	G2	G1	G2	G1	G2	
LighCNN9 (Categorical cross entropy loss)	87.36	82.46	88.96	82.71	77.24	81.88	90.88	84.98	2.708
LightCNN9 (Focal loss)	85.49	83.84	90.04	87.55	78.74	81.88	89.44	86.32	1.301
Linear SVM	83.36	83.05	90.47	75.11	73.95	86.61	88.96	82.30	4.374

Table 6.2. Gender prediction across all Age groups (A1, A2, A3, A4)

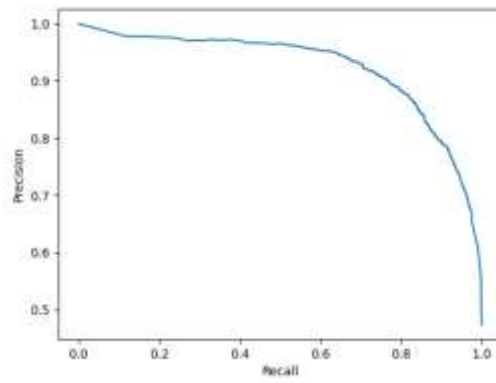
Model	Classification Accuracy (in %)								DoB (in %)
	A1		A2		A3		A4		
	G1	G2	G1	G2	G1	G2	G1	G2	

Analysis: We can visualize from *Table 6.1.* and *Table 6.2.* that LCNN performs better across both Demographic groups Age and Race as it has low Degree of Bias (DoB).

C. ROC Curve (Focal loss used):



D. Precision-Recall Curve (Focal loss used):



Question2: Bias and Explainability

Dataset Used: Cifar10

Dataset Split:

Split the dataset into:

1. **Train set:** 40,000 samples
2. **Validation set:** 10,000 samples
3. **Test set:** 10,000 samples

Algorithm: VGG16

- ### 1. Colored images for Test set:

A. Test Accuracy: 75.83 %

B. Confusion Matrix:

```
confusion matrix:
[[772  18  42  18  19   1  10  12  72  36]
 [  7 907   1   8   0   1   8   1  23  44]
 [ 83   4 642  61  71  45  50  17  13  14]
 [ 21   9  47 650  66  81  42  48  11  25]
 [ 20   3  42  63 739  23  34  65   4   7]
 [  3   4  34 214  48 601  14  62   6  14]
 [  6   7  25  64  41  15 810  17   6   9]
 [ 21   2  16  36  36  21   1 852   1  14]
 [ 38  15   8  16   3   2  12   2 879  25]
 [ 21  75   6  12   2   2   3   4  24 851]]
```

- ## 2. Gray scaled images for Test set:

A. Test Accuracy: 10.0 %

B. Confusion Matrix:

[illegible]

3. Class-wise Accuracy:

Class	Airplane (0)	Automobile (1)	Bird (2)	Cat (3)	Deer (4)	Dog (5)	Frog (6)	Horse (7)	Ship (8)	Truck (9)
Accuracy	80.53%	90.43%	56.70%	62.21%	73.90%	72.03%	74.12%	83.40%	86.85%	84.31%

4. Bias Metrics:

- Degree of Bias (DoB) = $std(CAcc_{Dj}), \forall j$
- Average False Rate (AFR) = $\frac{FPR+FNR}{2}$
- False Positive Rate (FPR) = $\frac{FP}{TN+FP}$
- False Negative Rate (FNR) = $\frac{FN}{TP+FN}$

VGG 16 model:

1. Degree Of Bias (DoB): 10.26 %
2. Average False Error Rate (AFR): 0.1343
3. FPR: (Averaged over each class): 0.027
4. FNR: (Averaged over each class): 0.2417
5. Consistency: There is not consistency in each class classification accuracy for classes (sub demographic groups) like bird and cat so model predictions are biased towards other classes except bird and cat classes.

Analysis for DoB: Yes there is very less Bias according to DoB is approximately 10% which implies there is less performance gap across different subgroups.

Analysis for AFR: (Performance Metric): There is very less bias as per the model performance error rate.

5. New Bias Metrics:

A. Ratio wise Disparate Impact for prediction for one Demographic group (Binary classification say Race prediction) and Bias evaluation across different Demographic group (having 2 subgroups say Gender1 and Gender2)

For Example:

As in the paper: **Subgroup Invariant Perturbation for Unbiased Pre-Trained Model Prediction**, [\(link\)](#) in table no.4 of this paper, we can see that for Race prediction Disparate impact is calculated across different demographic group Gender1 and Gender2 for Race1 and Race2 then average of these Disparate impact is taken but instead of that we can take ratio of Disparate impact across gender1 and Disparate impact across Gender2. It will give overall Disparate impact across all races.

B. Disparate impact for Multiclass classification problem: On being inspired through total probability formula used for Bayes theorem. We can calculate Disparate Impact (DI) for each class by calculating the probability across Sub Demographic Group1 divided by Total Probability across all Sub Demographic groups. Then bias factor for all classes can come in range of probability [0,1] and then we can compare bias in each class and mitigate bias according to each class.

Then we can move towards may be similar kind of approach for each type of present metric DoB, AFR and then can come up with updated PSE metric.

6. Bias Mitigation:

I. Data method (Pre - processing): Used Normalized data using Feature wise center and Feature wise standard normalization, and data augmentation for training data (rotation and horizontal flip).

- 1. Degree Of Bias (DoB):** 7.37 %
- 2. Average False Error Rate (AFR):** 0.112
- 3. FPR(Averaged over each class):** 0.022
- 4. FNR(Averaged over each class):** 0.202

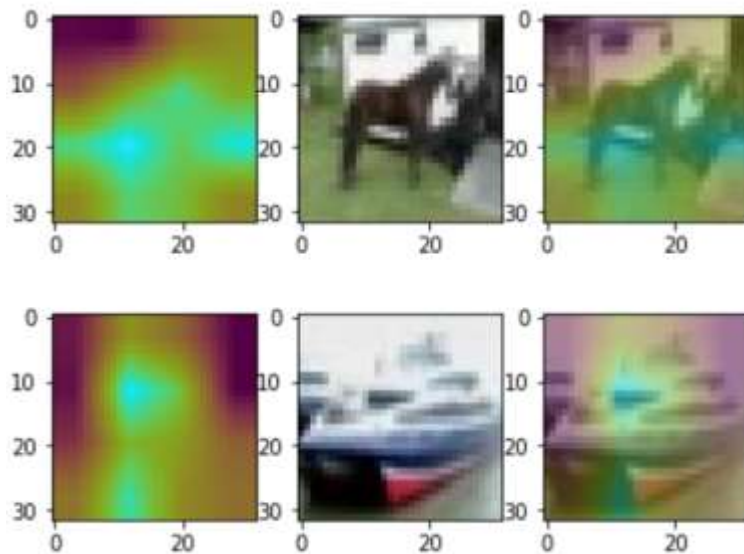
Analysis: As we can see DoB and AFR both has reduced hence we can say some Bias has been mitigated.

II. Algorithm method (loss function): Used KL Divergence loss function for this multiclass classification problem

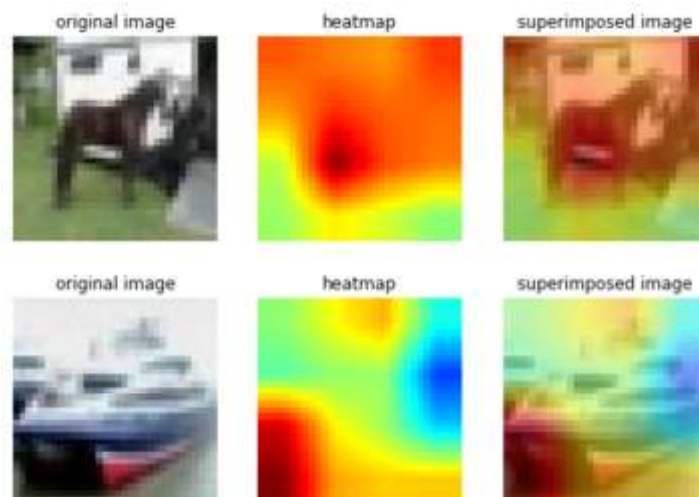
1. **Degree Of Bias (DoB):** 8.53%
2. **Average False Error Rate (AFR):** 0.12
3. **FPR: (Averaged over each class):** 0.024
4. **FNR(Averaged over each class):** 0.2231

Analysis: As we can see DoB and AFR both has reduced hence we can say some Bias has been mitigated.

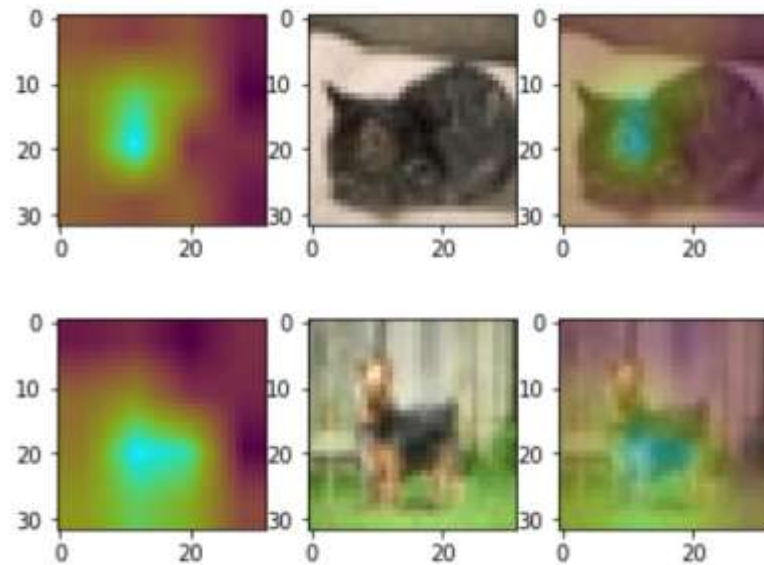
7. A. GradCam for Correctly Classified Samples (Visualized from layer : block4_conv2) :



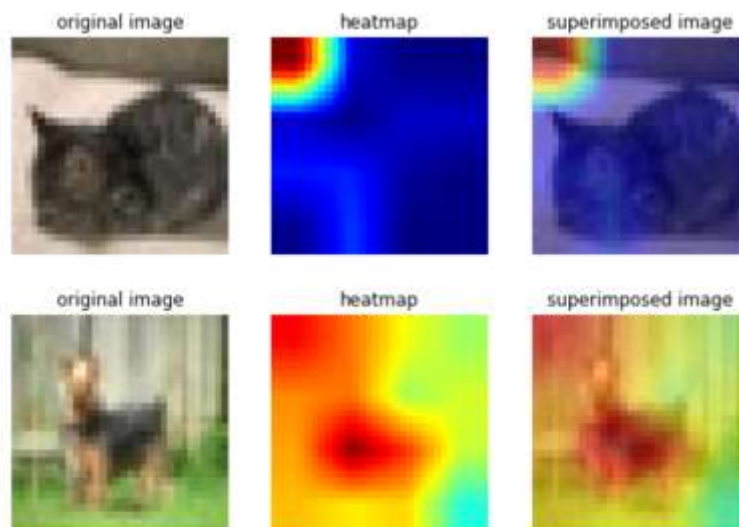
B. GradCam++ for Correctly Classified Samples(Visualized from layer : block4_conv2):



8. A. GradCam for Incorrectly Classified Samples(*Visualized from layer : block4_conv2*):



B. GradCam++ for Incorrectly Classified Samples(*Visualized from layer : block4_conv2*):



Grad-CAM Vs Grad-CAM++ Visual Analysis:

1. From 7A we can visualize for correctly classified object where more intense region (red color) of GradCAM heatmap focusing on, whereas in 7B we can visualize more obsolete focus on horse through GradCAM++. They are focusing on the portion which explains the correct classification prediction.
2. From 8A we can visualize for incorrectly classified object where more intense region (red color) of GradCAM heatmap focusing on other regions except target whereas in 8B we can visualize more focus on outside region instead of cat through GradCAM++. They are focusing on the portion which makes the incorrect classification prediction.
3. Clearer explanation for the corresponding action is shown in the case of GradCAM++ as compared to Grad-CAM.

References:

- [1]. https://kornia.readthedocs.io/en/latest/_modules/kornia/losses/focal.html
- [2]. https://github.com/samson6460/tf_keras_gradcamplusplus
- [3]. <https://github.com/AlfredXiangWu/LightCNN>
- [4]. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization
- [5]. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks