The VGG_MC_SL reports:

The Architecture of the NN is based VGG multi-class single label classifier neural network as it is explained previously. The test accuracy of the current classifier is around 80%.

The Precision, recall, Accuracy, and F-score are respectively defined as follows:

Precision: TP/ (TP+FP)

Recall: TP / (TP+FN):

Accuracy: (TP+TN)/total:

F-score: 2 * (Recall * Precision) / (Recall + Precision)

Where, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

Intuitive interpretation of these terms are as follows:

Accuracy, this term demonstrates how accurate the classifier is.

Precision explains how the classifier is good at **not predicting the false Positives**. While, the **recall** determines how the classifier is good at not making false negatives.

Let’s say we have a spam detector in this case, the **precision** has a higher priority rather than recall; that is, the lower values of False Positive, the higher precision __Less un-spam emails are detected incorrectly.

When we are dealing with a medical diagnosis, let’s say breast cancer detection, the **recall** has a higher priority. It is equivalent to say not detecting a sick person as a healthy person; to loss the opportunity of treatment definitely, has a higher priority.

To evaluate the performance of applying the threshold technique I performed the test in three cases. Before describing the test details let’s make some naming conventions.

Naming Conventions:

1) The term **others**’ refers to those images which are downloaded from SVPRESSA.RU. All of these images are consist of photos such that they do not contain any object which is used to train the classifier.
2) The **Classes** is referred to those classes which classifier is trained to make predictions and returns them as its output.

I design the test procedure as follows:

Case1- Equal: The number of images in others and classes are equal.

Case2- 2Others: The number of images in others is twice more than the number of photos in classes.

Case3- 2Classes: The number of images in classes is twice more than the number of photos in others.

Note: Per each class 6 images are selected, i.e. 7*6 = 42 images in classes, and 42 images in others. More details in the corresponding tables. Since we are dealing with multi-class classifier I utilized weighted average in all of the reported results.

In the Figure below test results for each of this cases are demonstrated.

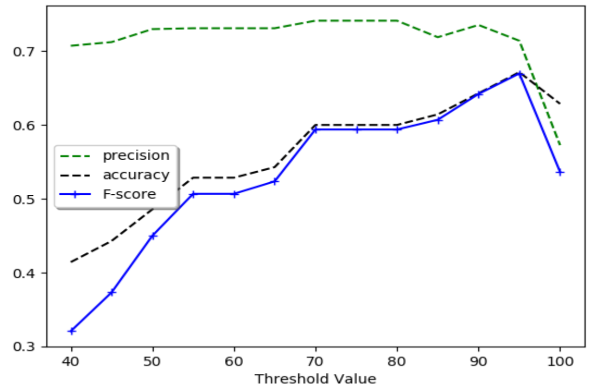

Figure 1: Equal

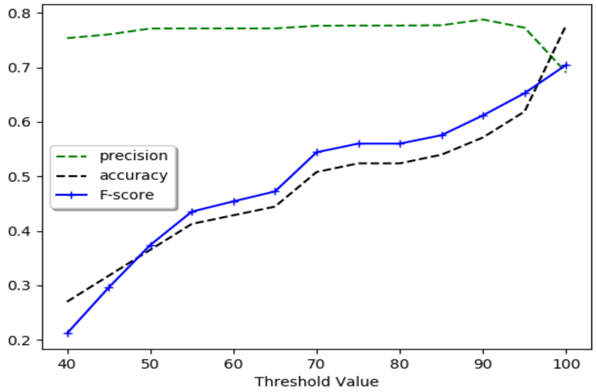

Figure 2: 2Others

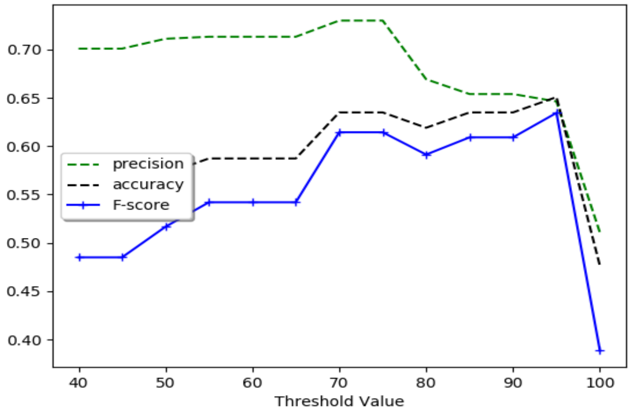

Figure 3: 2Classes

To interpret the figure and make conclusion I should define another term which I call it probability. **Probability is refers to, a criterion which measures how likely the input image of the classifier can be one of the classes of the classifer.** This probability can be defined based on some background knowledge, e.g. website content, website meta description, user knowledge, texts and ….

Figure 1 shows the case Equal. In this case, the precision decreases due to the increase of the threshold. especially, when it becomes very strict, that is, above the general accuracy of the classifier itself. Since half of the number of test samples do not contain any class of the classifier, thus the classifier produces some wrong outputs, obviously, accuracy and F-score are low for lower threshold and by increasing the threshold these terms are also increased.

Figure 3 represents the case 2, classes. The behavior of the classifier is almost the same to the previous case. Since in this case, the number of images in the OTHERS is half of the 2 CLASSES, the value of accuracy and F-score are accuracy are larger than the similar situation in Figure 1. The ripples in the figure may happen because of some overfitting.

Figure 2, justifying the behavior of the classifier is not hard just one should remind himself that the number of images in 2 OTHERs is 2 times more than the 2 CLASSES. Thus the accuracy and the F-score are almost half of the corresponding values in the 2 CLASSES case.

Note: However, I don't know the reason for the strange behavior of F-score and the accuracy, and the precision for the threshold between 95-100 in Figure 1, 3.

Base on above discussion I propose the following adaptive thresholding technique.

Adaptive Threshold:

Based on some background knowledge, we should set the threshold adaptively. To explain the framework we should define a term probability.

1) If the probability is high, let’s say the probability for an input image to be one of the classes of the classifier is more than two times of not being one those classes, we should set the threshold between the range 70 to 80.
2) If the Probability is low, let’s say the probability for an input image NOT to be one of the classes of the classifier is more than two times of being one those classes, we should set the threshold between the range above 90.

	No. Images in 2*classes	No. Images in 2*other	Precision						
			40	50	60	70	80	90	100
Equal	35	35	70	71	72	73	72	70	55
2*others	21	42	73	74	76	75	76	74	70
2*classes	42	21	70	71	71	72	65	63	51

	No. Images in 2*classes	No. Images in 2*other	F-score						
			40	50	60	70	80	90	100
Equal	35	35	32	47	50	55	61	65	65
2*others	21	42	21	29	46	55	56	60	68
2*classes	42	21	48	52	54	60	59	60	39

	No. Images in 2*classes	No. Images in 2*other	Accuracy						
			40	50	60	70	80	90	100
Equal	35	35	42	50	53	60	61	65	65
2*others	21	42	28	37	45	56	56	58	78
2*classes	42	21		57	58	63	62	63	47