Optimization for greedy non-maximum suppression based on multi-task convolutional neural network

Qingqing Hong\textsuperscript{a}, *, Qian Fan\textsuperscript{b}, Lifeng Zhang\textsuperscript{a}

\textsuperscript{a}Department of Electronical and Electronic Engineering Kyushu Institute of Technology
1-1, Sensui-cho, Kitakyushu, Fukuoka 804-8550, Japan
\textsuperscript{b}College of Information Engineering, Yangzhou University, Yangzhou, CHINA

* Corresponding Author: hongqingqing1221@gmail.com

Abstract

Non-maximum suppression (NMS) is an essential part of the face detection pipeline based on convolutional neural network (CNN). The common approach for NMS used by face detection is a greedy, locally optimal strategy which is to localize object from a set of candidate locations. However, NMS still has some shortcomings, such as sometimes the detection box has no relationship with high classification score, which leads to misjudge face localization during NMS. In this paper, we observed that NMS implemented on the multi-task convolutional neural network (MTCNN) which is a cascaded convolutional network and enhance NMS based on MTCNN to achieve high performance during face detection and alignment.

We employ WIDER FACE as test dataset to evaluate our proposal. The precision and recall curve is adopted with three sub sets at different thresholds. The result shows that our proposal can perform better performance than traditional NMS.

keywords: Face detection, face alignment, MTCNN, CNN, NMS.

1. Introduction

Face detection and alignment have been widely fundamental applications in computer vision which are to generate bounding boxes for identifying frontal face and assign them classification scores. However, the various visual of faces will impose great challenges for these tasks such as occlusions, large pose variation and extreme lightings in practical applications.

Besides the challenge of various faces’ vision, few exist approaches concentrated on the inherent correlation between face detection and alignment will also decrease the accuracy of face detector. Even if Chen et al.\textsuperscript{4} presented a new state-of-art approach for face detection and alignment. The key proposal is to adopt random forest which take the differences of pixel value’s differences into consideration. But the performance still limited by handcraft features. In this paper, we introduce a new framework represented by multi-task cascaded convolutional networks (MTCNN) to integrate face detection and alignment proposed by Kaipeng Zhang et al.\textsuperscript{5} The advantages of MTCNN is described as follows: First, MTCNN adopted a new cascade and lightweight CNN framework to combine face detection and alignment for achieving good real time performance. Second, an effective approach which is online hard sample mining was proposed in order to improve the performance. Because original hard sample mining could decrease the performance due to offline manner. The detail of MTCNN we will describe in the next chapter. Although, MTCNN’s result has performed better consistently compared to other state-of-the-art techniques like P. Viola et al.\textsuperscript{1}, H. Li et al.\textsuperscript{2}, Z. Zhang et al\textsuperscript{3} and Chen et al.\textsuperscript{4}.

However, there are still some shortcomings exist in MTCNN. After classification and regression scores are generated by MTCNN, there is no constraint that can delete redundant bounding boxes and generate specific box for one face. In order to avoid this problem, MTCNN employed greedy non-maximum suppression (NMS) as post-processing stage to obtain final detection boxes. But the feature of greedy NMS is to make hard decision and set a fixed threshold to decide the range of suppression. In this case that, face misjudged will be occurred and the average precision of entire network may be dropped.

To overcome the deficiency of greedy NMS, we will support a new improved NMS algorithm implemented on MTCNN. The improved NMS is to decay the scores of re-
dundant boxes that have an overlap with specific box as a consistent function.

The rest of this paper is organized as follow: Chapter 2 related work that describe the detail of MTCNN and NMS. Chapter 3 specify the exist issue of NMS remained in MTCNN. Chapter 4 describe our research’s proposal. Chapter 5 describe our experiment and result.

2. Related Work

In this chapter, we will describe the detail of MTCNN proposed by Kaipeng Zhang et al. and greedy NMS algorithm.

2.1 MTCNN

The overall frameworks of MTCNN is shown as Fig. 1. This framework performed a cascaded architecture which consists of three stages network.

When input one image into the framework, the framework will initially resize this image into different scales in order to establish a pyramid. And the result of pyramid will be regard as input for three-stage network that we will describe in the follows:

(a) Proposal Network (P-Net):

In this stage, a fully convolutional network was implemented. And adopted a similar approach proposed by S. S. Farfade et al. to gain the candidate windows and bounding box regression vectors. And then, estimated bounding box regression vectors was used to calibrate the candidates. The bounding box regression is formulated as follows by employing Euclidean loss ($y_{box}^{box}$ indicate the regression target obtained from the network and $y_{i}^{box}$ is the ground truth):

$$L_{i}^{box} = \| \hat{y}_{i}^{box} - y_{i}^{box} \|_2^2$$  (1)

The last step is to employ NMS to merge the high overlapped candidates. The architecture of P-Net is shown as Fig. 2. Since face detection is a binary classification task, less filters should be needed but more discrimination. In this case that, form Fig. 2 we can observe that the number of filters has been reduced and the size of filter has been changed to 3x3.

(b) Refine Network (R-Net):

The second stage of the MTCNN is R-Net. The basic architecture is convolutional neural network. Compared to the P-Net, a fully connected layer was added. After the input image processed by P-Net, many candidates windows still existed and then input to R-Net. Hence, large numbers of candidates will be rejected strictly by R-Net. The last step also the same as P-Net, calibrate candidates with bounding box regression and merge the candidates by using NMS. The architecture of R-Net is shown as Fig. 3.

(c) Output Network (O-Net)

The third stage of MTCNN is R-Net. Compared to the R-Net, the architecture of O-Net is more complexity. Since
Fig. 4. The architecture of O-Net

one more convolutional layer was added. The distinction between R-Net and O-Net is that the structure of O-Net will recognize the area of face through more supervision, and finally return five facial features. The architecture of O-Net is shown as Fig. 5.

Fig. 5. NMS removed the redundant candidate boxes.

2.2 NMS

NMS is frequently used as an essential part of face detection which is post-processing step for removing the redundant candidate windows and obtains impressive effect shown as Fig. 5. The NMS’s principle is that search for local maximums and suppress non-maximum values.

The workflow of NMS is described as follows: first, NMS starts with a list of detection boxes \( B \) which sorted by the confidence scores \( S \); Second, select the detection box which has maximum score \( M \) and add it into final detection set \( D \). Meanwhile, remove the box \( M \) which has the maximum from the list \( B \). Then we set a fixed threshold \( N_t \) and compute the remaining boxes’ intersection-over-union (IOU) with box \( M \) in the list \( B \). And then remove the box from the list \( B \) which has greater IOU than threshold \( N_t \). This process will be repeated for the remaining boxes in the list \( B \). We will take an example to explain the process more vividly. As Fig. 5 shows, there exist four bounding boxes \( A, B, C, D \). And sorted by confidence scores as \( A > B > C > D \). Then mark bounding box \( A \) as the maximum score box. Compute the IOU that remaining box \( B, C, D \) overlap with box \( A \). Make an assumption that \( D \)'s IOU greater than the threshold we set in advance. Then we will remove box \( D \) from the candidates. After that, we select the box which has higher confidence score from \( B \) and \( C \). And compute the IOU between \( B \) and

\[
\text{Input: } B = \{ b_1, \ldots, b_N \}, \quad S = \{ s_1, \ldots, s_N \}, \quad N_t
\]

\( B \) is the list of initial detection boxes 
\( S \) contains corresponding detection scores 
\( N_t \) is the NMS threshold

begin
\[
D \leftarrow \{ \}
\]
While \( B \neq \text{empty} \) do
\[
m \leftarrow \text{argmax } S
\]
\[
M \leftarrow b_m
\]
\[
D \leftarrow D \cup M; \quad B \leftarrow B - M
\]
For \( b_i \) in \( B \) do
\[
\text{If } \text{iou} (M, b_i) \geq N_t \text{ then}
\]
\[
B \leftarrow B - b_i; \quad S \leftarrow S - s_i
\]
end
end
return \( D, S \)

Fig. 6. Specific the process of NMS.

Fig. 7. NMS removed the redundant candidate boxes.
C. If IOU is greater than the threshold, the box which has the lower confidence score will be dropped.

The pseudocode also performed as following Fig. 7. Although NMS shows good performance during face detection. But due to its characteristics, NMS still has some limitations. We will describe it in the next section.

3. Pitfalls in traditional NMS

3.1 Pitfalls in traditional NMS

As we mentioned in the previous chapter, NMS shows good performance implemented in the modern face detection techniques. But some pitfalls still exist in the NMS which will cause miss and false detection occurred. The major pitfall of NMS presented as follows:

(a) NMS set the score for adjacent detection box to zero forcibly which means remove the detection box whose IOU is greater than the threshold. In this case that, if two or more objects are close to each other, all but one of the candidates will be removed, the result will be performed as

(b) NMS will maintain the box with highest confidence score. However, some cases show that the neighboring boxes with lower score may perform better for the true object like Fig. 8 (b) shows.

(c) NMS will return the candidate boxes which are not removed. Meanwhile, many boxes with relative lower confidence that haven’t detected the true object will also been returned. This problem will be shown as Fig. 8 (c) specifically.

Due to the pitfalls we described in the previous, the traditional NMS will degrade the performance of MTCNN which take NMS as last step in each stage to merge candidate boxes.

In next section, we will adopt a new NMS algorithm proposed by Navaneeth Bodla et al. for improving the performance of MTCNN consistently. The new NMS will decay the detection scores of all other neighboring boxed as a continuous function instead of setting the score to zero forcibly.

4. Proposal

Before we describe the new NMS, let’s review the traditional NMS with a simple function presented as equation (2):

\[
s_i = \begin{cases} 
    s_i, & \text{IOU}(M, b_i) < N_t \\
    0, & \text{IOU}(M, b_i) \geq N_t 
\end{cases}
\]

The parameter of M, \(b_i\), \(N_t\), and \(s_i\) have described in the previous chapter. Hence, we can simply find that NMS sets a fix threshold \(N_t\) which decide the candidate boxes which need to be removed or not. If some boxes’ IOU smaller than threshold, NMS will set the score to zero. To avoid the deficiency of traditional NMS, the new NMS adopts a linear function that decay the confidence score of neighbor boxes which have overlaps with the highest confidence score box M. In fact, we should consider that the boxes have higher overlaps with M needed to be decayed more. Hence, Navaneeth Bodla et al. proposed a equation presented as follows to achieve that the boxes which are far away from M will not be affected and boxes which are very close would be assigned a greater punishment.

\[
s_i = \begin{cases} 
    s_i, & \text{IOU}(M, b_i) < N_t \\
    s_i (1 - \text{IOU}(M, b_i)), & \text{IOU}(M, b_i) \geq N_t
\end{cases}
\]
From the Eq. 3, we should consider another problem that
the linear function is not continuous. When the IOU exceeds
the overlap threshold $N_t$, mutation will be occurred. In this
case that, a large changes of the detections’ rank list will be
happened. For taking this problem into consideration, another
function which can be continuously decay the confidence
score should be designed. The function should achieve the
following conditions:

(a) Boxes with lower IOU which means far away from M,
the penalty should be increased gradually.

(b) Boxes with higher overlaps with M, the penalty should
be increased significantly.

Based on the conditions above and mutation occurred in
previous linear function, Gaussian function has been adopted.
The formula presented as Eq. 4.

$$s_i = s_i e^{-\frac{\text{IOU}(M, b_i)^2}{\sigma}}, \forall b_i \notin D$$ \hspace{1cm} (4)

Where $\sigma$ represents the weight of Gaussian Function. Each
iteration should apply for this function and update the remain-
ing boxes’ score.

Next chapter, we will employ the improved NMS into
MTCNN to identify the improvement of original framework
used by traditional NMS. Our purpose is to achieve higher
average precision and shorten the entire detection time.

5. Experiment

5.1 Environment

We performed our experiments for identifying our proposal
that shows good performance compared to traditional NMS.
The WIDER FACE would be adopted to performed during
our experiment which is a face detection benchmark dataset.
The WIDER FACE choose 32, 203 images and label 393,
703 faces with a high degree of variability in scale, pose
and occlusion as depicted in the sample images. For each
event, 40%, 10% and 50% are selected as training, validation
and testing sets. In this work, we employed the validation
set as our test dataset. Because if we adopt the test dataset,
we need apply for official organization. Our experiment im-
plemented on Deeplearning-Box with GeForceGTX 1080Ti
11GB HDMI.

5.2 Measure metrics

Since we need compare the result with traditional NMS
implemented on MTCNN. We will perform the improvement
as recall and precision. The recall means the ratio of the
number correctly detected to the number of all faces in the
test set. The equation will be shown as Eq. 5.

$$\text{Recall} = \frac{tp}{fp + fn}$$ \hspace{1cm} (5)

Where the $tp$ represented by true positives and $tn$ repre-
sented by false negatives. And $n$ is the total images detected
by detectors. Denominator of eq.4 represent that how many
images with faces exist in the test set.

Another metric is precision. Precision represent that the
ratio of true positive in the detection images as Eq.6 showed:

$$\text{Precision} = \frac{tp}{tp + fn} = \frac{tp}{n}$$ \hspace{1cm} (6)

The $tp$ and $fn$ mean the true positives a false positives.
The sum of $fp$ and $tp$ is $n$ which represent total number of
images which has been detected. Based on these two metrics,
we specify the improvement of our proposal specifically.

5.3 Result analysis

In this section, we will describe our detail of experiment
and analysis our result compared to traditional algorithm. We
designed a program that can produce a texture file for each
detecting boxes. In this file, the format formed by image’s
name and number, the number of faces exist in the image,
and coordinate. Shown as Fig. 9. Than the result will be
computed with “wider face val.txt” which supported by offi-
cial organization of WIDER FACE to calculate the precision
and recall. The “wider face val.txt” is used to be recorded the
true positive information.

In our experiment, in order to show significant improve-
ment we utilized second stage of MTCNN which is R-Net to
evaluate our proposal. In the experiment, we set the weight
$\sigma$ of proposed gaussian NMS as 0.5. And the threshold $N_t$
for each stage of MTCNN as [0.5, 0.6, 0.6]. The result will
be shown below Fig. 10.

In Fig. 10, we show the PR curve and average precision
with three subsets: Easy, Hard, Medium. The hard set is
a super set of easy and medium which means contain all
faces taller than 10 pixels, the performance on hard set can
represent the performance on the entire test more accurately.
From Fig. 10, We specify that with the recall increased, the
precision will be decreased. Because if recall become higher,
the number of boxes will become more. In this case that
the precision will be decreased. Our result shows that our
proposal can achieve a high performance based on MTCNN.
Meanwhile, we set the different thresholds for calculating the
PR curve. From the result of PR curve, we can see among the
Fig. 9. Format of generated txt file

Fig. 10. Precision vs Recall produced by three functions based on MTCNN

three subsets with different thresholds, our proposed NMS can achieve better result than traditional NMS. The average precision can be improved 11.7% ～ 15.8%.

To take the result of Fig. 10 into consideration, we used two images selected from WIDER FACE dataset. That we can clarify that the detection boxes will be suppressed more, and the precision also increased. We will employ proposed NMS with gaussian function to compare to traditional NMS. The effect pictures are shown as Fig. 11.

6. Conclusions and future work

In this paper, in order to improve the performance of MTCNN, we adopted a new NMS algorithm proposed by Navaneeth Bodla et al. The proposed NMS is to decay the confidence score continuously instead of setting the score to zero forcibly which is the principle of traditional NMS. In the experiment, we utilized WIDER FACE dataset to identify the improvement of our proposal implemented on MTCNN. And calculate the recall and precision at different thresholds. The new NMS algorithm shows a good result compared to traditional NMS. From Fig. 11 and Fig. 11, the detection boxes are significantly decreased after applying new NMS and the average precision also achieved good improvement from 11.7% to 15.8%. However, from these two results, we can find that there still several faces which can’t be detected. And the average precision on hard set is a bit worse, since proposed NMS based on MTCNN focused on learning easy faces.

In this case that, we will take further research hard faces and improve the precision. And we will also make more effort to improve the performance of MTCNN through decreasing the number of faces which can’t be detected.

References

(4) D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun: "Joint cascade face detection and alignment", European Conference on Computer Vision, Volume, pp. 109-122, 2014
(6) S. S. Farfade, M. J. Saberian, and L. J. Li: "Multi-view face detection using deep convolutional neural


(16) B.Yang, J. Yan, Z.Lei, and S. Z. Li: “Convolutional channel features”, in IEEE International Conference on Computer Vision, 2015, pp. 82-90