1 USING NETWORK TO PREDICT PUPIL CENTER DIRECTLY

There are 2 reviewer mention that the capability of predicting pupil center directly for network:

> **Reviewer 1:** *(4) The networks which are used, VGG-16 and AlexNet have 138 and 61 millions of parameters, respectively. My major point is that these networks are capable of predicting pupil center, themselves.*
>
> **Reviewer 4:** *4. suggestions for revisions.- Please report the comparison with baselines with the CNN-based methods and also directly using VGG network for pupil center regression.*

As soon as we receiving the comments, we perform the experiment of using network to predict pupil center directly although we tried shallow CNN before.

1.1.1 Experiment 1: Deep CNN

The parameters settings that have been tested:

Architecture	VGG(modified)
Loss function	Mean Square Error
Learning rate	10^{-3} $for\ without\ pretrained\ weights$
Pre-trained weights	$10^{-3}, 10^{-4} \ldots, 10^{-8}$ $for\ initialed\ with\ pretrained\ weights$

Explanations

1. Architecture modification: we change the output layer to 2 and delete the softmax layer which means we use the 2 output scalars to represent the predicted coordinate. The classification model is convert to an regression model.
2. Pre-trained weights: as it's widely used in fine-tune model, we load layers exclude last 3 fully connected layers.
3. Other details can be found in the training code.

Unfortunately, although we explore many parameters settings including whether use pre-trained model, whether use shuffle batch and different learning rate, the loss change sharping during the training process and finally become nan.

We found there are some discussions about whether deep network is capable of regression. And it seems neural network work fine when the network is shallow, but when network gets deeper the network become unstable mainly due to the loss and gradient's problem, which explain the result of our experiment called loss concussion.

1.1.2 Experiment 2: Shallow CNN

In fact, another student[1] of our group tried this idea and report the result to us at the beginning of our research in pupil tracking field. And the experiment is using shallow CNN to regress the coordinate of the pupil center, but the result is that the predicted coordinate is always "close to" the ground truth but hard to reduce the pixel error into [0,5] pixel (the resolution of the whole image is 640*480). According to his report, the shallow CNN can promise 90% frame has the pixel error [0,25] pixel, but has much lower accuracy when required pixel error is <= 5 pixel. The suspicious cause is the loss function, which is the Euclid distance between predicted point and ground truth, has too much tolerance to the samples that has >=5 pixel error. Our target is to reduce the number of frame that has >= pixel error while the loss function defined the target of learning algorithm which is to reduce the mean square error of training set. We suspect that it's this subtle differences cause the result that accuracy of the shallow CNN is high when pixel error tolerance is high (>=25) but drop sharply when pixel error tolerance reduce to [0,5].

1.1.3 Inspiration to the proposed method

After we found CNN's disadvantage, which is having difficulty in regressing accurate coordinate, we choose to use CNN as a classifier instead of regressor.

1.1.4 Conclusion

- From experiment 1, deep CNN architecture is not suitable for regression without thoroughly modifications in architecture just like CNN based method in object tracking field. The reason is that the loss and gradient concussion, and network become unstable.
- From experiment 2, Shallow CNN had hard time regressing precise coordinates (5 pixel error in 640*480 image) although it's capable of getting rough coordinates (>=25 pixel error in 640*480 image).

From our experiments and researches, we don't think using network to regress the pupil center is suitable with current deep learning technique. Maybe when deep learning performs better in regression problem in the future, using network to regress the pupil center might work.

---

[1] The student, in fact is a high school student, left the group after we discussed this and he didn't participate the rest of the research, so he is not the author list. If it's not suitable, please let us know.