GAZEATTENTIONNET: GAZE ESTIMATION WITH ATTENTIONS

Haoxian Huang[1] *Luqian Ren*[2] *Zhuo Yang*[3]

School of Computers, Guangdong University of Technology, Guangzhou

ABSTRACT

Predicting the gaze point on mobile devices in unconstrained environments without calibration has a great significance on human computer interactions. Appearance-based methods for gaze estimation have significantly improved due to the recent advances in convolutional neural network (CNN) models and the availability of large-scale datasets. But these models may have limitations on extracting the global information of features and may ignore the information of spatial features. In this paper, we propose a novel structure named GazeAttentionNet. We use the global and local attention module to utilize both global features and local features comprehensively improving the accuracy of gaze estimation. Firstly, we use MobileNetV2 and the self-attention layers as the global attention module to extract global features. Secondly, we add the local attention module containing the spatial weight module to extract local features. With this GazeAttentionNet structure, we achieve a good result on the GazeCapture dataset. The average error of mobile phones and tablets is 1.67 cm and 2.37 cm with an improvement of 18% and 28% compared to the original experiment result on the GazeCapture dataset.

Index Terms— gaze estimation, self-attention, spatial weight mechanism

1. INTRODUCTION

Gaze is an integral part of humans in the visual world. The way how people interact with the environment is mainly through the gaze. We can acquire information from our surroundings while expressing our potential thoughts through the gaze. Recently, gaze estimation has been widely used in many scientific research fields, including human computer interaction [1], assisted driving [2], psychology [3], and so on. Also, with the development of hardware, recent gaze estimation researches focused on utilizing commodity hardware like webcams or the front-facing cameras available in ubiquitous mobile phones and tablet PC devices. These devices can capture people's appearance images which can be used as inputs of the appearance-based models of the gaze estimation directly. Besides, these widespread devices can bring great convenience to our daily lives with applications based on gaze interaction.

The methods of gaze estimation can be divided into model-based and appearance-based [4]. The model-based approaches are firstly used to address the gaze estimation problem by using geometry features of eyes. Lately, the availability of large datasets and novel deep learning technologies make appearance-based methods possible to have great performance on gaze estimation. The CNN models are widely used on image classification, object detection, and other computer vision tasks due to their dramatic performance of image processing. With the great ability of feature extraction, CNN models are also general models used in gaze estimation tasks. For example, Krafka et al. [5] proposed the iTracker model which used the convolutional neural network as a basic model to extract features of the left eye, right eye, and faces images respectively. These features will be combined with the face grid input to estimate the gaze point on mobile devices by using the fully connected layers. But most of the CNN models may not have great use of global features and spatial features limiting them to achieve better performance on gaze estimation.

Our contributions are as follows:

(1) We combine MobileNetV2 and the self-attention layers as a global attention module.

(2) We add the local attention module with the global attention module proposing the GazeAttentionNet structure.

(3) We can achieve a superior result on the GazeCapture dataset with an average error of 1.67 cm and 2.37 cm on the mobile phones and tablets respectively.

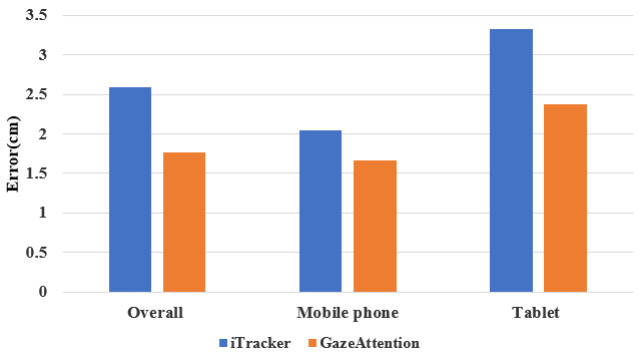

Fig. 1. The performance of iTracker and the GazeAttentionNet model on the GazeCapture dataset.

2. RELATED WORKS

2.1. Model-based Approaches

Model-based methods explore the characteristics of human eyes to identify a set of distinctive features around the eyes. The limbus, pupil, and corneal reflections are common features used for eye localization. Then model-based methods utilize these visual features to fit a geometric 3D eye model to perform gaze estimation. Model-based methods can be subdivided into corneral reflection and shape-based methods, depending on whether they rely on external light sources to detect eye features. Valenti et al. [6] used the shape-based methods to estimate the gaze direction combining the features of eyes and the head pose. These traditional approaches tend to suffer from low image quality and variable lighting conditions.

2.2. Appearence-based Approaches

Appearance-based approaches directly use the images of eyes or face as the inputs of the model mapping the image data to the gaze vector or gaze point. Although the appearance-based methods can achieve a satisfying result of gaze estimation, these methods need a large number of data to train the model. Krafka et al. [5] introduced the large-scale dataset of mobile gaze estimation named GazeCapture containing data from over 1450 people and consisting of almost 2.5M frames. MPIIGaze [7] is a dataset for unconstrained 3D gaze direction estimation containing a large number of images from different participants. These images were collected from the participants' serval months of their daily life meaning the background of these images are different and the light condition also various. With these large-scale appearance-based datasets, we can also solve the many tasks related to the eye movements. In [8] Chang et al. proposed a highly efficient high-frame-rate eye-tracking method suitable for the performance-constrained mobile environment. In [9] Jha et al. proposed an approach converting the regression problem into a classification problem, predicting the probability at the output instead of a single direction to improve the accuracy of the gaze estimation task.

Appearance-based approaches have the potential to work on the low-quality images captured by webcams or front-facing cameras on the phones or tablets. Given the success of previous appearance-based gaze estimation researches and available huge labeled datasets, in this work, we focus on this kind of method. It's well known that the Transformer models which mainly consist of self-attention layers have great performance on processing the NLP tasks with its ability to utilize the global information of the sentences. In the paper, we combine the CNN models and the self-attention layers to acquire the global information of the features, and we add the spatial weight module to utilize the local features of the inputs to get a better result for the gaze estimation task.

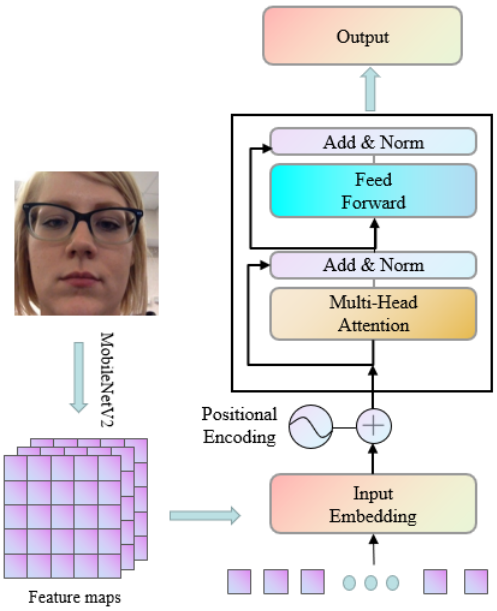

Fig. 2. The structure of the global attention module consists of the MobileNetV2 and the self-attention layers. The mobileNetV2 extracts the image's feature forming a 32×7×7 feature map. The feature map will be embedded with a positional encoder as the input of the self-attention layers.

3. GAZEATTENTIONNET

3.1. Global Attention Module

The structure of the global attention module is shown in Fig.2. The global attention module is made up of MobileNetV2 and the self-attention layers. MobileNetV2 is used to extract a compact feature representation of the image. And the self-attention layers are used to extract the whole information of inputs by computing a weighted sum of features maps. The self-attention layers are organized based on the structure of Transformer's encoder [10].

Backbone. MobileNetV2 is used to process the input images into feature maps. It contains the inverted residual module with linear bottleneck. This module takes as an input a low-dimensional compressed representation which is first expanded to high dimension and filtered with a lightweight depthwise convolution. So it can alleviate the problem of feature degradation, reduce the amount of computation required for convolution, and make use of the input's information better at the same time.

Self-attention Layers. The self-attention mechanism is an attention mechanism relating different positions of a single sequence to compute a representation of the whole sequence. Given a feature matrix X, the feature is projected into queries Q, keys K, and the values V. We calculate the scaled dot product of vectors Q, K, V and apply the softmax function to ob-

tain the weights on values. The operation of the self-attention mechanism can be summarized as [10]:

$$Attention(Q, K, V) = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d_m}})V \qquad (1)$$

where $\frac{1}{\sqrt{d_m}}$ is the scaling factor, controling the fluctuation of dot product.

In this paper, the self-attention layers is organized based on the structure of the Transformer's encoder composed of a stack of six identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a position-wise fully connected feed-forward network. And a residual connection around each of the two sub-layers, followed by layer normalization. Also, the fixed positional encodings are added to the input of each attention layer.

3.2. Local Attention Module

Attention mechanism was first used in natural language processing. At present, CNN based attention mechanisms are widely used in computer vision tasks. In [11] they proposed the CBAM module which applies attention to both channel and spatial dimensions to focus on the useful information in the image classification and object detection tasks. Also, attention mechanisms have been used in gaze estimation tasks. In [12] they first used the spatial weights module in the convolutional neural network to estimate the gaze direction using the face images as input and achieved an excellent result on the MPIIGaze dataset [7].

Spatial Weights Module. In this paper, the spatial weights module includes three convolution layers with filter size 1×1 followed by a linear unit layer. The input of the spatial weights module is an activation tensor U of size N×H×W, and the output is a H×W spatial weight matrix W. We use four convolution layers and the spatial weight module as the local attetion module. The structure of the local attention model is shown in Fig.3 .

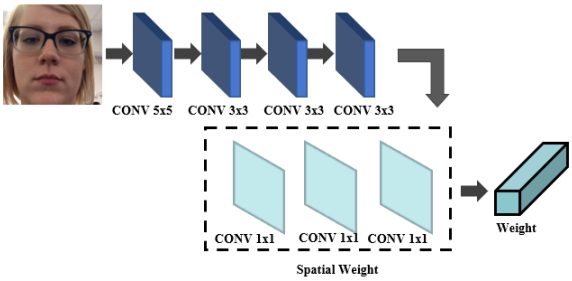

Fig. 3. The followings of the basic convolution layers in the first row are the ReLU function and Maxpool layer. In the spatial weight module, the ReLU function is added after the first two convolution layers separately. A Sigmoid function is added after a fully connected layer which is used to compress the size of the weight matrix.

3.3. Feature Fusion

As shown in Fig. 4, the proposed GazeAttentionNet structure consists of three components, including the global attention module, local attention module, and fully connected layers. A two-branch network is designed to extract the global and local information of the eyes and face images respectively. These feature maps will be fused after these two attention modules.

The left eye, right eye, and face images are resized to 3×224×224. Then, all of these images will be fed into the global and local attention module separately. We use the element-wise multiplication to fuse the output of the first and the second branch forming feature maps of the left eye, right eye, and face. The left and right eye's feature maps are combined by a fully connected layer as eyes' feature maps.

Lastly, we use the fully connected layers to combine the feature maps of the eyes and face with the face gird input outputting the gaze point.

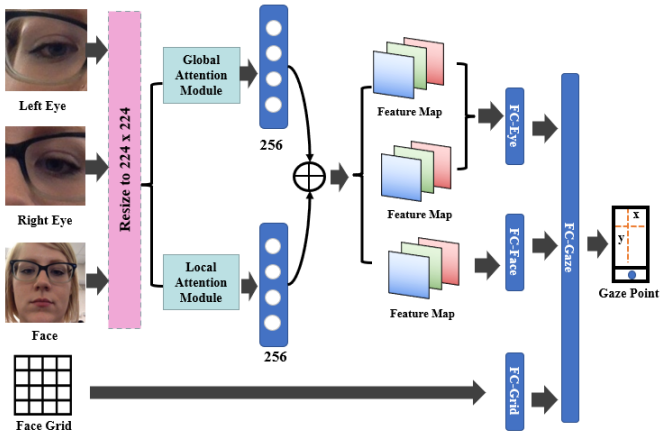

Fig. 4. The GazeAttentionNet contains the global attention module, local attention module, and fully connected layers. The global and local attention module is the first branch and second branch separately in Fig. 4. The details of these two modules are shown in Fig. 2 and Fig. 3.

4. EXPERIMENTS

4.1. Data Preparation

In this paper, we adopt the GazeCapture [5] dataset to test the performance of our GazeAttentionNet model. The GazeCapture dataset is a large-scale dataset containing almost 2.5M frames. All of the frames were collected by the front-facing cameras of the mobile devices. Besides, during the process of collecting data, the participants were asked to rotate the devices to change the position of the camera including putting it on the top, bottom, left, and right. These operations can make the dataset's data more diversified, which is beneficial to test the model's robustness. For training, each of the samples is

treated independently while for testing, we average the predictions of the samples to obtain the prediction on the original test sample.

4.2. Training Details

We implement our experiment and design by using PyTorch (version=1.9.0) and Python (version=3.8). To compare the performance of different models, the hyperparameters are set as the same in the overall experiments. The optimizer we used is the SGD optimizer with a learning rate of 0.0001, a momentum of 0.9, and a weight decay of 0.0005 throughout the training procedure. The batch size of the experiments is set to 32.

4.3. Preliminary Experiments

Before testing our model on the GazeCapture dataset, we have used the CNN models such as AlexNet, ResNet18, VGG11, GoogleNet, and mobileNetV2. The structures of these CNN models are different. For example, the AlexNet consists of convolution layers, pooling layers, and fully connected layers using a linear combination while the ResNet18 has a residual structure to reuse the features of former layers. All of them are the famous CNN models used in image classification with excellent performance on the ImageNet dataset.

These CNN models will be used to replace the convolution layers of the iTracker structure as the basic structure using eyes, face images, and face grid as inputs to test their performance on the GazeCapture dataset. The results of these experiments are shown in the top half of Table 1. We can find out that mobileNetV2 achieves an average error with the value of 1.75 cm and 2.67 cm on mobile phones and tablets, which is better than other CNN models. So we adopt MobileNetV2 as the backbone of our global attention module.

Table 1. Results of different CNN models and the GazeAttentionNet on the GazeCapture.

Model	Mobile Phone Error(cm)	Tablet Error(cm)
iTracker	2.04	3.32
AlexNet	1.85	2.83
ResNet18	1.82	2.72
VGG11	1.82	2.72
GoogleNet	1.79	2.73
MobileNetV2	1.75	2.67
GazeAttentionNet	**1.67**	**2.37**

4.4. GazeAttetionNet Experiments

4.4.1. Details of global attention module

Firstly, the image is processed by MobileNetV2 into a feature map with a size of 1280×7×7. It mainly consists of the deep separate convolutions and the inverted residual modules. Then, a convolution layer with a kernel size of 1×1 is added to scale the channel dimension of this high-level feature map. Finally, we can a 32×7×7 feature map containing the main information of the image.

Secondly, the input of the self-attention layers is a sequence. So we collapse the spatial dimensions of the feature map into one dimension, resulting in a 32×49 feature map with fixed positional encoders. Then, we feed the feature maps into the self-attention layers which are organized based on the Transformer's encoder containing eight heads self-attention mechanism. Through these operations, we can the global information of the whole image.

4.4.2. Details of local attention module

The image will be fed into four convolution layers with different kernel size to compress the image's information into a 256×28×28 feature map. Then, the spatial weigth module containing three convolution layers with a kernel size of 1×1 is used to get a weight of the feature map helping the model focus on the important parts of the image.

4.4.3. Experiments Results

At first, we test the global attention module's performance on the GazeCapture. We use it to replace the convolution layers of the iTracker. And the inputs are as same as the preliminary experiments. In the same experiment environment, we achieve an average error of with the value of 1.68 cm and 2.51 cm on mobile phones and tablets proving the self-attention layers can improve the performance. Then, we test the GazeAttentionNet's performance. Finally, we achieve the result shown in the last row in Table 1, which is better than other models'.

5. CONCLUSION

In this paper, we propose the GazeAttentionNet structure using the global and local attention modules to address the gaze point estimation task on mobile devices. GazeAttentionNet is capable of acquiring robust both global and local features. With the global attention module, we can get the global information of the features without splitting the images, which may cause losses of the important parts of images. Furthermore, the local attention module can help us to focus on the spatial features of inputs improving the utilization of the input images. Finally, with the GazeAttentionNet model, we can achieve a superior result on the GazeCapture.

6. REFERENCES

[1] Vijay Rajanna and Tracy Hammond, "Gawschi: Gaze-augmented, wearable-supplemented computer-human interaction," New York, NY, USA, 2016, Association for Computing Machinery.

[2] Congcong Liu, Yuying Chen, Lei Tai, Haoyang Ye, Ming Liu, and Bertram E. Shi, "A gaze model improves autonomous driving," in *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications, ETRA 2019, Denver , CO, USA, June 25-28, 2019*, Krzysztof Krejtz and Bonita Sharif, Eds. 2019, pp. 33:1–33:5, ACM.

[3] Nina A. Gehrer, Michael Schönenberg, Andrew T. Duchowski, and Krzysztof Krejtz, "Implementing innovative gaze analytic methods in clinical psychology: A study on eye movements in antisocial violent offenders," New York, NY, USA, 2018, Association for Computing Machinery.

[4] Dan Witzner Hansen and Qiang Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 478–500, 2010.

[5] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba, "Eye tracking for everyone," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2176–2184.

[6] Roberto Valenti, Nicu Sebe, and Theo Gevers, "Combining head pose and eye location information for gaze estimation," *IEEE Transactions on Image Processing*, vol. 21, no. 2, pp. 802–815, 2012.

[7] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 1, pp. 162–175, 2019.

[8] Yuhu Chang, Changyang He, Yingying Zhao, Tun Lu, and Ning Gu, "A high-frame-rate eye-tracking framework for mobile devices," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 1445–1449.

[9] Sumit Jha and Carlos Busso, "Estimation of gaze region using two dimensional probabilistic maps constructed using convolutional neural networks," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3792–3796.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. 2017, vol. 30, Curran Associates, Inc.

[11] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, Eds., Cham, 2018, pp. 3–19, Springer International Publishing.

[12] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 2299–2308.