

GAZEATTENTION: GAZE ESTIMATION WITH ATTENTIONS

Haoxian Huang¹ Luqian Ren² Zhuo Yang³

School of Computers, Guangdong University of Technology, Guangzhou

ABSTRACT

Gaze estimation is one of the most important fields of computer vision. Through gaze, we can analyze the state of a person finding out whether he is focused, and find out the place where he is looking at. With the development of deep learning, the accuracy of gaze estimation has been improved dramatically. In the previous works, most of the appearance-based models use the convolutional neural network (CNN) models as the basic network to predict the gaze direction. But these models may have limitations on extracting the global relationship of the features and may ignore the information of the spatial features. In this paper, we use the global and local attention module to utilize the global features and the local features comprehensively improving the accuracy of gaze estimation. Firstly, we use the MobileNetV2 and the self-attention layers as the global attention module to extract the global features. Secondly, we add the local attention module to extract the local features. With this GazeAttention structure, we achieve a dramatic result on the GazeCapture dataset. The average error of the iPhone is 1.67 cm and the iPad's average error is 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 attention

1. INTRODUCTION

Eye 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 the input of the appearance-based model of the gaze estimation directly. Besides, these common devices can bring convenience to our daily life 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 the geometry features of the 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 the common models used in gaze estimation tasks. For example, Krafka et al. [5] proposed the iTracker model which used the convolutional network as the basic model to extract the 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 by using the fully connected layers. But most of the CNN models may not have great use of the global features and the spatial features limiting them to achieve better performance on gaze estimation.

In this paper, firstly we combine the MobileNetV2 and the self-attention layers as the global attention module. Secondly, we add the local attention module with the global attention module proposing the GazeAttention model to get a better accuracy of gaze estimation on the GazeCapture dataset. The performance of GazeAttention on the GazeCapture dataset is shown in Fig.1 .

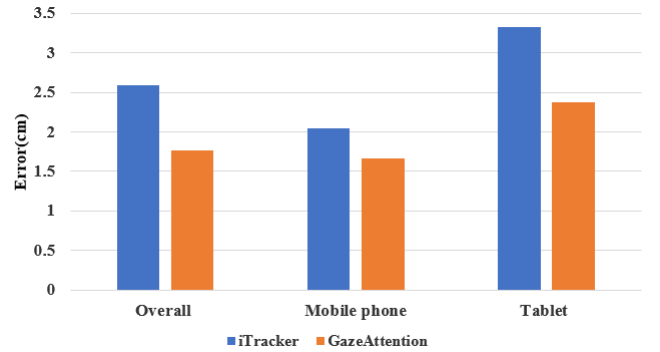


Fig. 1. The performance of iTracker and the GazeAttention model on the GazeCapture dataset.

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 corneal 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. Appearance-based Approaches

Appearance-based approaches directly use the images of eyes or face as the input 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. Also based on this dataset, they put forward an end-to-end method using eyes, face, and face grid as the input of the CNN model to estimate the gaze point on the mobile devices. 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' several months of their daily life meaning the background of these images are different and the light condition also various. Gaze360 [8] contains 172,000 images collected in indoor and outdoor environments with a wide range of head poses and distances between subjects and cameras. The varieties of the images can help the appearance-based models achieve a more robust performance on gaze estimation.

Appearance-based approaches have the potential to work on the low-quality images captured by the webcams or the front-facing cameras on the phone or the tablet. 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 relationship of the features, and we add the spatial attention module to utilize the local features of the input to get a better result for the gaze estimation task.

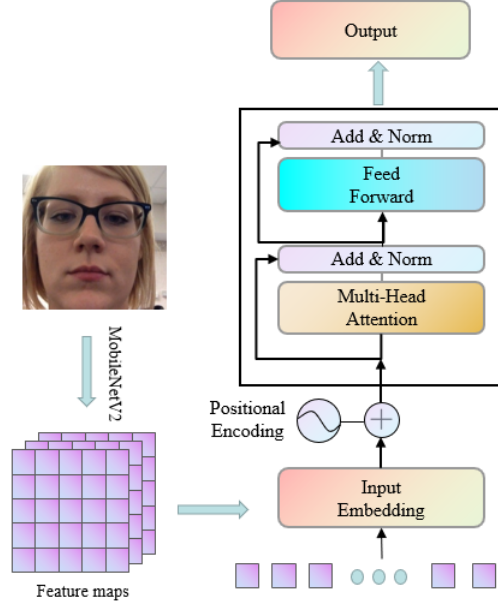


Fig. 2. The structure of the global attention module consists of the MobileNetV2 and the self-attention layers. The mobileNetV2 extracts the image's feature forming a $32 \times 7 \times 7$ feature map. The feature map will be embedded with a positional encoder as the input of the self-attention layers.

3. GAZEATTENTION

3.1. Global Attention module

Self-attention is an attention mechanism relating different positions of a single sequence to compute a representation of the sequence. In the self-attention layers, all of the keys, values, and queries come from the same place. Each position in the self-attention layers can attend to all positions in the previous layer. The Transformer [9] is a model architecture eschewing recurrence and instead relying entirely on the self-attention layers to draw global dependencies between input and output. The Transformer consists of self-attention layers based on the structure of the encoder and decoder. With its great ability to extract the global information of the inputs, Transformer models have achieved a great performance on the NLP tasks such as machine translation, speech recognition, and so on.

Also, people try to use the Transformer models in computer vision tasks. ViT [10] firstly used the pure Transformer model for the image recognition task. ViT splits an image into patches and provides the sequence of linear embeddings of these patches as an input to the Transformer. DERT [11] is a hybrid structure of CNN and the Transformer dealing with the object detection task. DERT contains a CNN backbone used to extract a compact feature representation, an encoder-decoder Transformer, and a simple feed-forward network that makes the final detection prediction.

Inspired by these, we adopt the MobileNetV2 network as the CNN backbone and combine the self-attention layers as the global attention module. We utilize the mobileNetV2 to extract the features of the images. At the same time, we use the self-attention layers with the multi-head attention structure to acquire the global information of the features without splitting the images. The self-attention layers map a query and a set of key-value pairs to an output, where the query, keys, values, and the output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by the dot-product function of the query with the corresponding key. Considering of the gaze estimation task is a regression problem, not a classification problem, we drop the decoder of the Transformer which is mainly used for the classification. The structure of the global attention module showing in Fig. 2 .

3.2. Local Attention Module

Humans tend to selectively concentrate on a part of information when and where it is needed, but ignore other perceivable information at the same time. Attention mechanism can be used as a resource allocation scheme to solve the problem of information overload. Inspired by the attention mechanism of humans, researchers bring the attention mechanism to the computer vision (CV) tasks and the natural language processing (NLP) tasks.

Spatial attention allows the neural networks to learn the positions that should be focused on. Through spatial attention, the spatial information in the original picture is transformed into another space and the key information is retained. Jaderberg et al. [12] introduced a different spatial transformer network (STN) that can find out the areas that need to be paid attention to in the feature map through transformations such as cropping, translation, rotation, scale, and skew.

Considering of the operations of cropping and rotation on the eyes and the face images may make a negative effect on the gaze estimation task, we extract the local features by using the spatial weights module. In [13] they used the spatial weights module in the convolutional neural network to estimate the gaze direction using the face images as input and achieved a great result on the MPIIGaze dataset [7]. The spatial weights module includes three convolutional 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 $N \times H \times W$, and the output is a $H \times W$ spatial weight matrix W . In this paper, we use four convolutional layers to extract the features of the images. After the convolutional layers, the feature maps will be fed into the spatial weight module containing three convolution layers with the kernel size of one. The structure of the local attention model is shown in Fig.3 .

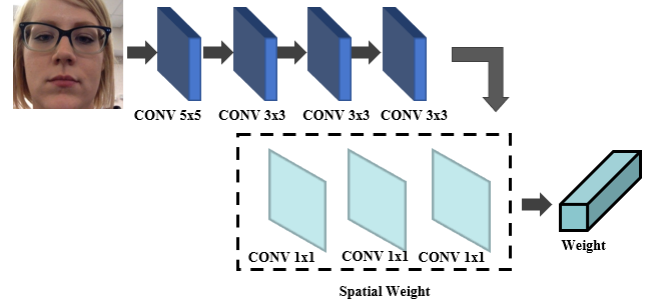


Fig. 3. The followings of the basic convolution layers in the first row are the ReLU function and the Maxpool layer. In the spatial weight module, the ReLU function is added after the first two convolutional layers separately and the Sigmoid function is added after the last convolutional layer.

3.3. GazeAttention

As shown in Fig. 4, the proposed GazeAttention consists of three components, including the global attention module, local attention module, and the fully connected layer for the face grid. The left eye, right eye, and face images are resized to $3 \times 224 \times 224$. Then, a two-branch network is designed to extract the features of the eyes and face images respectively, so that both the global and the local features can be obtained. In the first branch, we utilize the global attention module to learn the global features of the input images. In the second branch, we use the four convolution layers and the spatial weight module to get the local features of the images. We use the element-wise operation to fusion the output of the first branch and the second branch forming the feature maps of eyes and the face. Lastly, we use the fully connected layers to combine the feature maps with the face grid input outputting the gaze point.

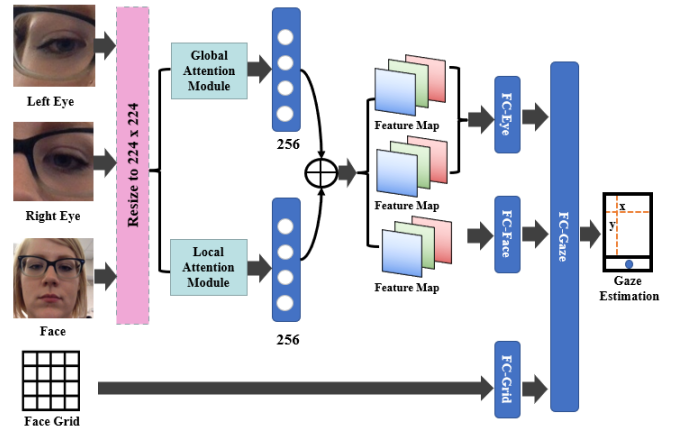


Fig. 4. The GazeAttention contains the global attention module and local attention module. The detail of the global and local attention module is shown in Fig.2 and Fig.3 separately.

4. EXPERIMENTS

4.1. Data Preparation

In this paper, we adapt the GazeCapture [5] dataset to test the performance of our GazeAttention model. The GazeCapture dataset is a large-scale dataset containing almost 2.5M frames totally. 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 common CNN models such as ResNet18, ResNet50, VGG16, GoogleNet, and mobileNetV2 as the basic model with the input of eyes, face image, and the face grid to test their performance. The results of these common CNN models are shown in Table 1 including the average error of mobile phones and tablets. We can find out that mobileNetV2 achieves a minimum error with the value of 1.94 and 2.5 on mobile phones and tablets. So we use the mobileNetV2 as the basic structure of our model.

Table 1. Results of the common CNN models on the GazeCapture dataset.

Model	Mobile Phone Error(cm)	Tablet Error(cm)
AlexNet	2.30	
ResNet18	2.12	
ResNet50	2.23	
VGG16	2.08	
GoogleNet	2.16	
MobileNet V2	1.94	

Table 2. Result of the GazeAttention model

Model	Mobile Phone Error(cm)	Tablet Error(cm)
iTracker	2.04	3.32
iTracker(no eyes)	2.11	3.40
iTracker(no face)	2.15	3.45
iTracker(no fg.)	2.23	3.90
GazeAttention	1.67	2.37
GazeAttention(no eyes)	1.84	2.69
GazeAttention (no face)	1.67	2.45
GazeAttention (no fg.)	0	0

4.4. GazeAttention Experiments

The global attention module is the first structure we utilized to process the input images. We can gain a feature map with the size of $1280 \times 7 \times 7$ after the convolutional processing of the MobileNetV2 model. After that, we use an extra 1×1 convolution layer to scale the channel and get $32 \times 7 \times 7$ feature maps. We feed the feature maps into a six layers transformer. As for the transformer, we set the hidden size of the two-layer MLP as 512 and perform eight heads self-attention mechanism. The dropout probability is set as 0.1.

We extract the local features by using the local attention module. The inputs of the local attention module are the same as the global attention module. The local attention module consists of 5 convolutional layers and the spatial weight module with 3 convolutional layers using the kernel size of 1×1 . With this structure, we can gain the spatial weights of the feature map guiding the model to focus on the useful features. The performance of the GazeAttention is shown in Fig. 4.

4.5. Ablation Experiments

To test our model's performance further, we use the eyes, face as the inputs of the GazeAttention respectively with the face grid and without the face grid. From the result shown in Table 2, we can acquire that the GazeAttention has a better performance compared to the iTracker in all experiments.

5. CONCLUSION

In this paper, we propose GazeAttention using the global and local attention module to address the gaze point estimation task on mobile devices. GazeAttention is capable of acquiring robust both global and local features, which can solve the problems of head position variation well. With the global attention module, we can get the global relationship of the feature without splitting the images, which may cause the loss of the important part of the images. Furthermore, the local attention module can help us to focus on the spatial feature of the input improving the utilization of the input image. The local

attention module can filter the irrelevant information making the model more robust.

Finally, with the GazeAttention model, we can achieve the mobile phone error and the iPad error with the value of 1.67, and 2.37 respectively with an improvement of 18% and 28% compared to the iTracker’s result.

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] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba, “Gaze360: Physically unconstrained gaze estimation in the wild,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6911–6920.
- [9] 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.
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *International Conference on Learning Representations*, 2021.
- [11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko, “End-to-end object detection with transformers,” in *Computer Vision – ECCV 2020*, Cham, 2020, pp. 213–229, Springer International Publishing.
- [12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu, “Spatial transformer networks,” in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, Curran Associates, Inc.
- [13] 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.