GAZEATTENTION: GAZE ESTIMATION WITH ATTENTIONS

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

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 local features and the global features comprehensively improving the accuracy of gaze estimation. Firstly, we use the MobileNetV2 and the self-attention layers to extract the global features. Secondly, we add the spatial attention module to extract the local features. With this 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 iTracker model.

Index Terms— Gaze estimation, self-attention, spatial attention

1. INTRODUCTION

Gaze has great meaning in our daily life. The way how people acquire the environment's information is mainly through the gaze. At the same time, we can also learn about the person's potential thoughts by analyzing the gaze. Recently, gaze estimation has been widely used in many scientific research fields, including human-computer interaction [2], assisted driving, psychology, advertising recommender systems 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. 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 []. 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 [1] used the CNN models as the basic model to estimate the gaze point's position using the appearance images as input. 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 module to improve the performance of gaze estimation on the GazeCapture dataset. Secondly, we add the local attention module with the global module named the GazeAttention model to get a better accuracy of gaze estimation. The structure of the GazeAttention model is shown in Fig.3. We find that using the local attention module and combing the global attention module can achieve great performance on the GazeCapture dataset. This approach can make the unconstrained gaze estimation based on mobile applications become possible.

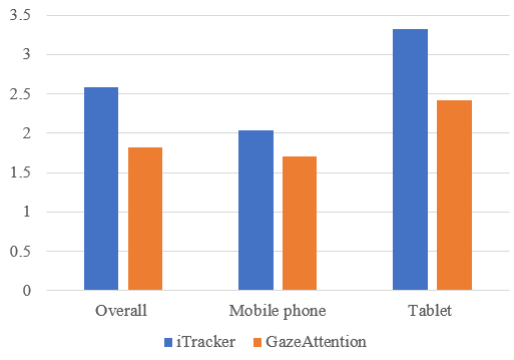

Fig. 1. The performance of iTracker and the GazeAttention model on the GazeCapture dataset.

2. RELATED WORK

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. [3] 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 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 [1] firstly 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 [4] 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. Gaze360 [5] 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. Because 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 improve the accuracy of gaze estimation. Also, we add the spatial attention module to utilize the local features of the eyes to get a better result for this task.

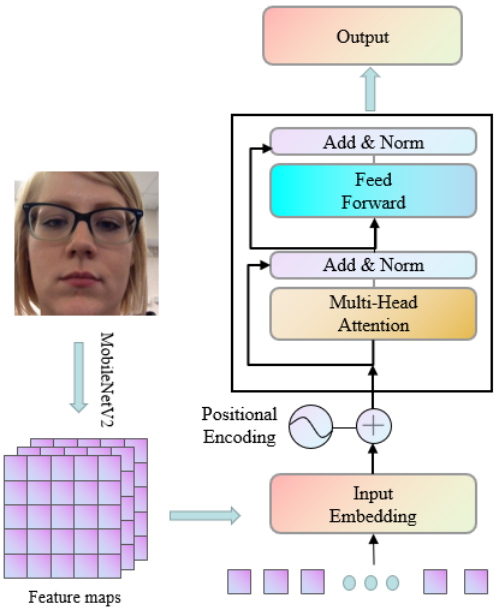

Fig. 2. The structure of the global attention module consists of the MobileNetV2 and the encoder of Transformer containing the self-attention layers. The mobileNetV2 extracts the image's feature forming a 32x5x5 feature map. The feature map will be embedded with a positional encoder as the input of the self-attention layers.

3. 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. Transformer [6] 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. The encoder and the decoder architectures have been used in many attention mechanism-based models such as the recurrent neural network (RNN). 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 [7] 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 pathes as an input to the Transformer. DERT [8] is a model based on the Transformer dealing with the object detection task. DERT is a hybrid structure of CNN and the Transformer. A CNN backbone to extract a compact fea-

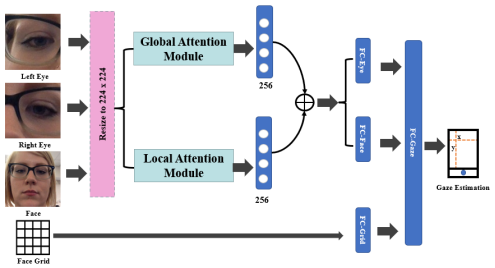

Fig. 3. The eyes and the face images are resized to 224 x 224 and fed into the global attention module and local attention module separately. We use the element-wise operation to fusion the output of the global attention module and the local attention module. Lastly, the feature of eyes, face, face grid will be combined by the linear layers maps the gaze point on the mobile devices.

ture representation and 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 to estimate the gaze point. We want to utilize the mobileNetV2 to extract the features of the images. At the same time, we use the self-attention layers to acquire the global information of the features without splitting the images. 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. Table 1 shows the results of the common CNN models and Transformer models using the eyes, face, and the face grid as inputs as same as the iTracker [1] model.

4. 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. Mnih et al. [9] used the attention mechanism on the recurrent neural network model to classify images. Bahdanau et al. [10] used the attention mechanism to simultaneously perform translation and alignment on machine translation tasks.

Spatial attention allows 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. [11] 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 mechanism. In [12] they used the spatial weights mechanism 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 [4]. The spatial weights mechanism includes three convolutional layers with filter size 1x1 followed by a linear unit layer. The input of the spatial weights mechanism is an activation tensor U of size N x H x W, and the output is a H x W spatial weight matrix W. In this paper, we use four convolutional layers to extract the feature of the images. After the convolutional layers, the feature map will be fed into the spatial weight module containing three convolutional layers with the kernel size of one. The structure of the local attention model is shown in Fig.4. We extract the local features using the spatial weights module and extract the global relationship of the feature maps with the global attention module proposing a model named GazeAttention.

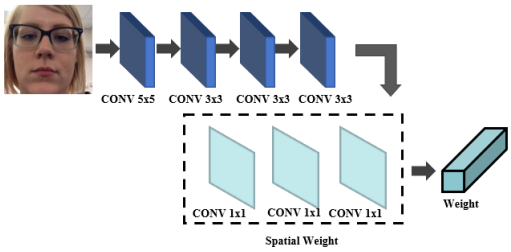

Fig. 4. The followings of the basic convolution layer in the first row are the ReLU function and the Maxpool layer. In the spatial weight module, the ReLU function is added after the first two convolutional layers separately and the Sigmoid function is added after the last convolutional layer.

5. EXPERIMENTS

In this paper, we adapt the GazeCapture [1] dataset to test the performance of our GazeAttention model. The GazeCap-

ture 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 the dataset's data more diversified and more efficient to test the model's robustness.

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 and a momentum of 0.9.

5.1. Baseline Experiments

Before using our model as the basic model of the iTracker structure on the GazeCapture dataset, we have used the common CNN models such as ResNet18, ResNet50, VGG16, GoogleNet, and mobileNetV2 as the basic model of the iTracker to test their performance. The results of these common CNN models are shown in Table 1. From Table 1, we can find out that mobileNetV2 achieves a minimum error with the value of 1.94 on the whole dataset. So we use the mobileNetV2 as the basic structure of our model.

Table 1. Benchmark of the common CNN models and Transformer models

Network Model	Error(cm)
AlexNet	2.30
ResNet18	2.12
ResNet50	2.23
VGG16	2.08
GoogleNet	2.16
MobileNet V2	**1.94**
ViT	2.51
LeViT	2.02
CVT	**1.94**

5.2. GazeAttetion Experiments

The global attention module is the first structure we used to experiment. We can gain a feature map with the size of 7x7x1280 after the convolutional processing of the MobileNetV2 model. After that, we use an extra 1x1 convolutional layer to scale the channel and get 7x7x32 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 1x1. With this structure, we can gain the spatial weights of the feature map guiding the model to focus on the useful features.

Table 2. Result of the experiment

model	Mobile phone error	Tablet error
iTracker	2.04	3.32
iTracker(no face)	2.15	3.45
Global attention module	**1.68**	2.51
Global attention module (no face)	1.71	2.49
GazeAttention	1.70	**2.42**
GazeAttention (no face)	0	0

5.3. Ablation Experiments

To test our model's performance further, we use the eyes, face as the inputs of the GazeAttention model respectively with the face grid and without the face grid. From the result shown in Table 2, we can acquire that GazeAttention without face image also can achieve a better result than the iTracker model using the eyes, face, and the face grid as the input.

6. CONCLUSION

In this paper, we propose a global and local attention network GazeAttention 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 overall error, mobile phone error, and the iPad error with the value of 1.82, 1.70, and 2.49 respectively compared to the iTracker's 2.59, 2.04, and 3.32.

7. REFERENCES

[1] 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.

[2] Tommy Strandvall, "Eye tracking in human-computer interaction and usability research," in *Human-Computer Interaction – INTERACT 2009*, Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler, Eds., Berlin, Heidelberg, 2009, pp. 936–937, Springer Berlin Heidelberg.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., Cham, 2020, pp. 213–229, Springer International Publishing.

[9] Volodymyr Mnih, Nicolas Heess, Alex Graves, and koray kavukcuoglu, "Recurrent models of visual attention," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds. 2014, vol. 27, Curran Associates, Inc.

[10] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, 2014.

[11] 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.

[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.