Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

An Overview of the Attention Mechanisms in Computer Vision

To cite this article: Xiao Yang 2020 *J. Phys.: Conf. Ser.* **1693** 012173

An Overview of the Attention Mechanisms in Computer Vision

Xiao Yang

The College of Electronic Science and Technology, National University of Defense Technology, Changsha 430070, China

yangxiaogfkd@163.com

Abstract. Deep convolutional neural network (CNN) plays an important role in the field of computer vision and image processing. In order to further improve the performance of CNN, scholars have conducted a series of new explorations, such as the improvement of activation functions, the construction of new loss functions, the regularization of parameters and the development of new network structures. However, every breakthrough of CNN comes from the innovation of network structure, whose design can be inspired by exploring the cognitive process of human brain. As one of the important features of human visual system, visual attention mechanism is essential in image generation, scene classification, target detection and tracking when applied in the field of computer vision. Focusing on the models of attention mechanisms commonly used in computer vision, their categorizations, principles, and outlook are summarized in this overview.

1. Introduction

Attention mechanisms originated from the investigations of human vision. In cognitive science, only a part of all visible information is noticed by human beings due to the bottlenecks of information processing. Inspired by this visual attention mechanism, researchers have tried to find the model of visual selective attention to simulate the visual perception process of human beings, so as to model the distribution of human attention when observing images as well as videos and expand its applications[1]. In recent years, important breakthroughs of attention mechanisms have been made in the fields of image and natural language processing. It has been proven that attention mechanisms can improve the performance of models, and they are also consistent with the perceptual mechanism of human brain and eyes. Taking the field of computer vision for example, most research combining deep learning and visual attention mechanisms concentrates on the use of mask. The principle of mask is that the key features in the image data are identified by another layer with new weight. By learning and training, deep neural network can learn the areas where attention needs to be paid in each new image, thereby forming attention. This idea further evolved into two different types of attention: soft attention and hard attention. The mechanism of soft attention is realized via gradient descent and is of differentiability and continuity. In neural networks, the weight of attention can be learned through forward propagation and backward feedback [2]. Hard attention mechanism, however, is not differentiative, which is often achieved by reinforced learning and motivated by the benefit function to make the model pay more attention to the details of some parts.

This paper will introduce in three parts: the first part is the computational models of visual selective attention; the second part is the classification of the attention mechanism models of computer vision; the third part is the summary of the existing attention mechanisms and an outlook.

2. Computational Models of Visual Selective Attention

Visual selective attention is of great importance in the mechanism of human visual system, since it can guide us to grasp the important contents within the scene and enable the visual system to obtain useful information with limited processing resources. Although this activity is subjective, it is undeniable that the nature of contents has an influence on selective attention, which reflects the top-down and bottom-up attention processing mechanisms of visual selective attention computation.

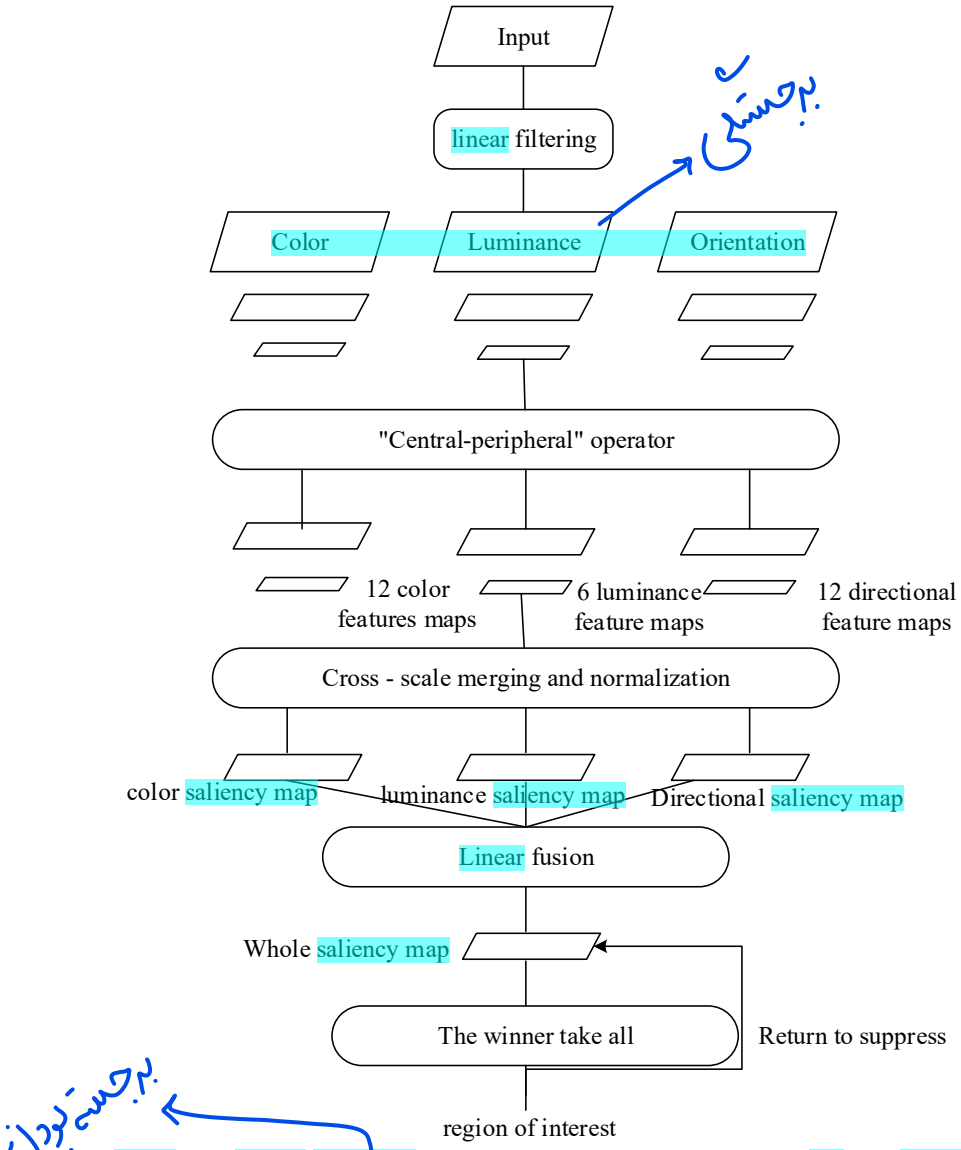

Fig.1 The visual attention computation model proposed by Itti and Koch

At present, the research on the modeling of visual selective attention is mainly about four aspects: how to compute the bottom-up saliency map; how to describe and express top-down tasks and knowledge; how to merge the bottom-up information flow with the top-down knowledge flow; how to control the shift of attention focus [3]. Based on biological principles, Itti and Koch proposed a bottom-up visual attention computing model, which effectively simulates the human visual selective attention mechanism and is currently the most widespread model [4]. As evident in Fig.1, this model is divided into two parts: the computation of saliency map as well as the selection and transfer of attention region. The core of saliency map computation is the extraction and fusion strategy of the features of each channel, which is also the core of the whole model framework. Firstly, the input

image is decomposed into three multi-channel image components: color, luminance and orientation, respectively, and a multi-resolution pyramid is established to represent each component. Subsequently, these multi-scale component images are operated by a "center-periphery" operator, which simulates the characteristics of human sensory field, so 12 color feature maps, 6 luminance feature maps as well as 24 orientation feature maps are obtained. Afterwards, corresponding color, luminance and orientation saliency maps can be attained by normalization linear overlapping, and a total saliency map can be produced based on these three saliency maps of different channels. Finally, a manually built dynamic neural network selects the attention region via the saliency map. The selected region is a circle centered with the focus of attention, and points with larger values in the saliency map will be noticed first.

3. The classification of attention mechanism models in computer vision

By applying the bottom-up visual attention computing models to the specific tasks of computer vision, the interference of irrelevant information can be ignored, so the generalization performance of network can be improved. In the process of development, attention mechanisms in computer vision have evolved into different categories, and different models pay attention to diverse feature domains. In this section, these models will be introduced by typical examples.

3.1. Soft attention

Due to the differentiability of soft attention, it has been used in many fields of computer vision, such as classification, detection, segmentation, model generation, video processing, etc. Mechanisms of soft attention can be categorized into spatial attention, channel attention, mixed attention, self-attention.

3.1.1. Spatial attention

Ordinary CNN can show the translation-invariance and implicit rotation-invariance of learning. Compared with the networks learning things implicitly, an explicit processing module is preferred for the network to handle all the abovementioned transformations. Consequently, DeepMind designed Spatial Transformer Layer (STL) to realize spatial invariance [5], and its network structure is demonstrated in Fig.2.

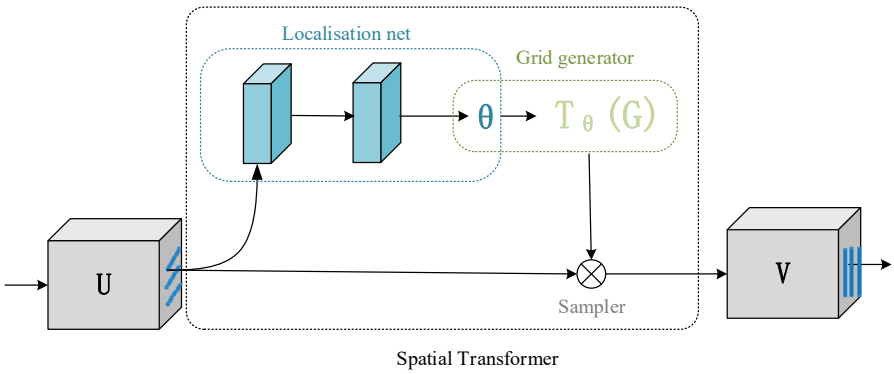

Fig. 2 Structure of the STL model

The localization net firstly obtains a Θ according to the input image, U, by computation. And the grid generator then computes the coordinates of input image according to Θ and the coordinates of output image. In the end, the sampler fills image V based on the defined rules of filling (bilinear interpolation is generally used). In this case, the input image can be corrected into the desired image by STL.

3.1.2. Channel attention

In a convolutional neural network, each image is initially represented by three channels (R, G, B). After being processed by different convolution kernels, each channel will generate new channels containing different information. If weights are added to each channel to show the relevance between channel and key information, a greater weight means a higher relevancy, and more attention should be paid to the corresponding channel.

SENet, the winner of ImageNet Classification Contest in 2017, is essentially a channel-based attention model [6]. It models the importance of each feature channel and then enhances or suppresses it in different tasks. The principle is displayed in Fig.3.

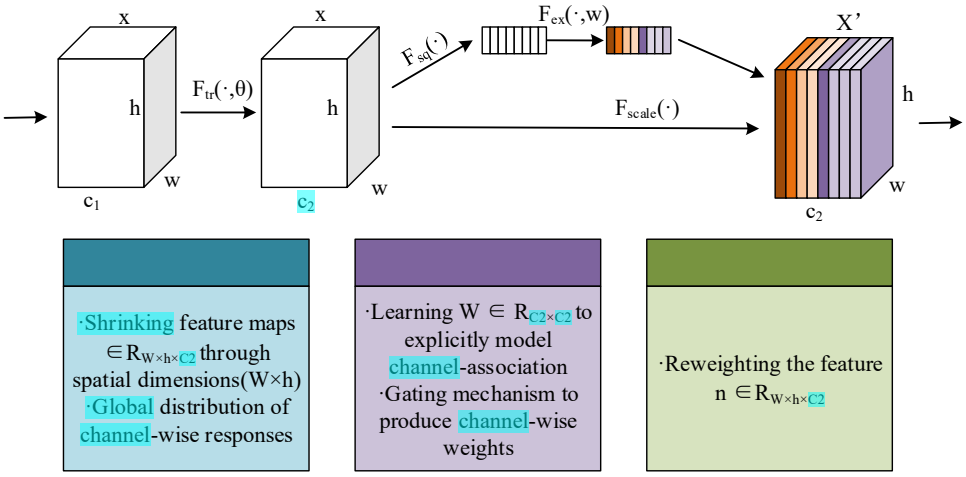

Fig. 3 Principle of SENet

A by-pass branch emerging after normal convolution is operated by squeezing, which compresses the features of spatial dimension, that is, each two-dimensional feature map becomes a real number. The next step is the operation of excitation, which generates a weight w for each feature channel to explicitly model the relevance. Once the weight of each feature channels is obtained, the weights are applied to each original feature channel, and the importance of different channels can be learned according to specific tasks.

When applied to several benchmark models, this mechanism can achieve a significant improvement of the performance though increasing a small amount of computations. As a general idea of design, it can be utilized in many existing networks. Afterwards, such methods as SKNet can enhance the performance by combining the idea of channel weighting and the multi-branch network structure in inception. The essence of channel attention mechanism lies in the modeling of importance among various features, whose weights can be assigned according to the input in various tasks, therefore, this mechanism is simple and effective.

3.1.3. Mixed attention

The mixed attention combines multiple attention mechanisms into one framework, which can bring better performance to a certain extent. CBAM firstly combined the mechanism of channel attention and spatial attention [7], making the network know not only 'look what' but also 'look where' by just adding a few more parameters. The network structure of CBAM is demonstrated in Fig. 4.

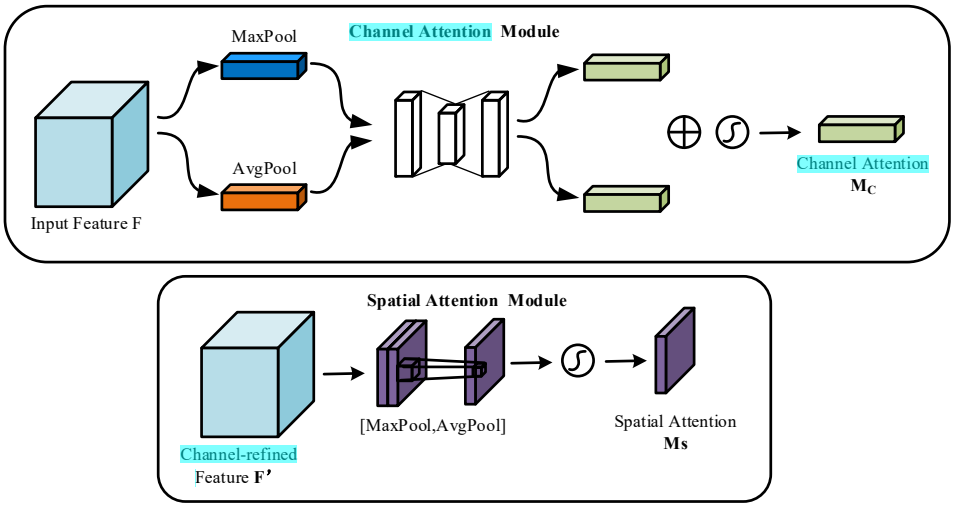

Fig.4 Structure of CBAM

3.1.4. Self attention

In a convolutional neural network, the convolution kernel is confined by its size, which can only use local information to calculate the target pixel, so it may lead to deviations due to the ignorance of global information. If each pixel in the feature map is regarded as a random variable and the paring covariances are calculated, the value of each predicted pixel can be enhanced or weakened based on its similarity to other pixels in the image. The mechanism of employing similar pixels in training and prediction and ignoring dissimilar pixels is called the self-attention mechanism.

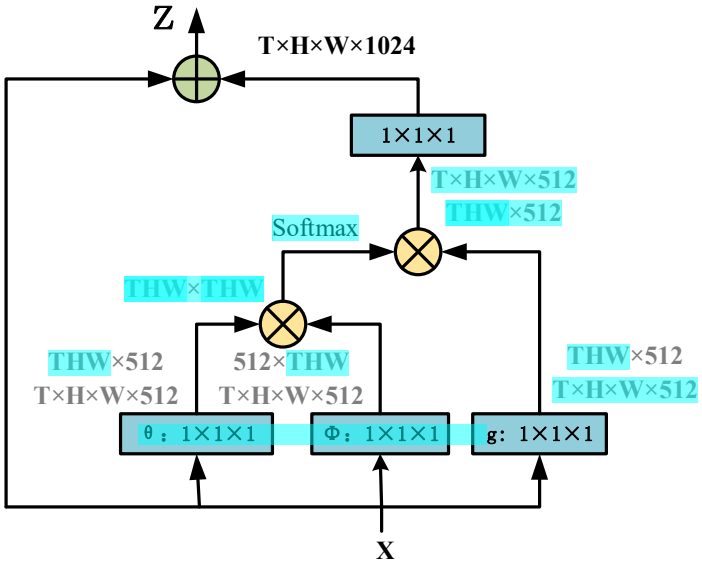

Fig.5 Structure of Non-local Neural Networks

To achieve global reference for each pixel-level prediction, Wang et al. put forward the non-local Neural Networks of self-attention in CNN, as shown in Fig.5 [8]. Their approach is to treat each pixel as a random variable based on the predicted covariance between pixels. The participating target pixels are the weighted sum of all pixel values, where the weights are the correlation between each pixel and the target pixel. By using the self-attention mechanism, global reference can be realized during the training and prediction of models. The model is with good bias-variance weight, making it more reasonable.

3.2. Hard attention

The mechanism of soft attention has been widely and successfully applied in the field of computer vision. However, the research on the mechanism of hard attention in computer vision tasks is relatively limited. As the mechanism of hard attention can select important features from input information, it is regarded as a more efficient and direct approach. Although the role of constraints such as sparsity in shaping the ability of learning agents has been explored, AttentionAgent took a different way and was inspired by concepts related to inattentional blindness, that is, when the brain is engaged in a task requiring effort, it pays most attention to the elements related to the task, and temporarily ignores the other signals.

To realize this, [9] segments the input image into several blocks, and then simulates the voting among blocks based on the modified self-attention architecture, thereby selecting a subset that is considered important. Relevant blocks are selected in each time step, and once determined, AttentionAgent will make decisions only according to these blocks while ignoring the other blocks. Usually, back propagation is utilized to optimize neural networks. However, considering that AttentionAgent includes non-differentiable operations, such as sorting and slicing, to generate important blocks, it is not easy to apply such techniques to training. Therefore, the algorithm of non-derivative optimization is adopted to overcome this problem, as shown in Fig.6.

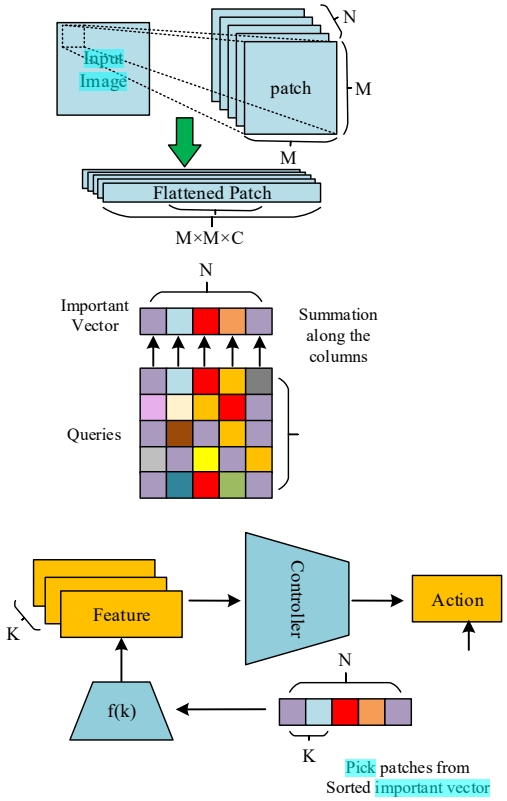

Fig. 6 Principle of [9]

The upper row: input transforming - the sliding window splits the input image into smaller blocks and then “flattens” them for future processing. The middle row: block election - the modified self-attention modules vote between blocks to generate a vector of block importance. The lower row - action generation - AttentionAgent selects the most important blocks, extracts corresponding features, and makes decisions on their basis.

It has been proven that AttentionAgent has successfully learned to pay attention to different regions in the input image. The visualization of key blocks allows to peer into how agents make decisions, thereby demonstrating that most choices are significant and intuitive to humans, making it a powerful

tool for analyzing and debugging agents in development. Besides, as the agents have also learned to ignore information that is not important to the core task, they can be generalized to tasks in which the environment has been slightly modified.

4. Conclusion

Based on the overview provided above, the conclusions are obtained as below:

(1) Attention mechanism is an intersection of computer science, biology, psychology and cognitive science, so the cognitive rules of human beings need to be understood in the first place. Until now, many cognitive characteristics of human vision are still to be explored, and there is no unified theoretical framework for reference in the field of human visual attention mechanisms. Only by further strengthening the research of human attention mechanisms, exploring and modeling the rules of human vision in information processing, can it be better applied to the field of computer information processing.

(2) The combination of low-level and high-level visual features makes us pay different attention to varied things, and these features also contribute differently. Consequently, it is of essence to choose appropriate weighting methods to make it more in line with the patterns of human eyes in observing things.

(3) The process of human visual attention is realized through the combination of bottom-up processing of primary visual features and top-down task guidance. However, most models are processed in a single way. In this case, the combination of bottom-up and top-down computing models can better simulate the mechanism of human visual attention, which has a broader research prospect in the field machine vision.

(4) Due to the addition of the time axis feature, videos are more complex than natural images, and there are more limited computing models for video visual attention. Under this circumstance, describing the objects in salient regions of videos with semantics of higher dimensions can make it closer to human vision, which is beneficial to fields such as the retrieving and classification of videos.

References

[1] Feng Hui. On the Mechanisms and Application of Visual Attention Mechanisms [D]. North China Electric Power University (Beijing).
[2] Zhao B , Wu X , Feng J , et al. Diversified Visual Attention Networks for Fine-Grained Object Classification[J]. IEEE Transactions on Multimedia, 2017, 19(6):1245-1256.
[3] Liu Qiong, Qin Shiyin, Li Zhicheng. The Modeling Computation of Visual Selective Attention and Its Application Prospect [J]. Science & Technology Review, 2010(01):107-115.
[4] Itti L , Koch C , Niebur E . A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 20(11):1254-1259.
[5] Jaderberg M , Simonyan K , Zisserman A , et al. Spatial Transformer Networks[J]. 2015.
[6] Hu J , Shen L , Albanie S , et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8):2011-2023.
[7] Woo S , Park J , Lee J Y , et al. CBAM: Convolutional Block Attention Module[J]. 2018.
[8] Wang X , Girshick R , Gupta A , et al. Non-local Neural Networks[J]. 2017.
[9] Tang Y , Nguyen D , Ha D . Neuroevolution of self-interpretable agents[J]. 2020.