A Temporal-Spatial Attention Model for Medical Image Detection

Maxwell Hwang[1], Cai-Wu[2], Kao-Shing Hwang[3],
Yong Si Xu[3], Chien-Hsing Wu[3]
[1]Department of Colorectal Surgery, the Second Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang, China,
[2]Department of Hematology, the Fourth Affiliated Hospital of Zhejiang University School of Medicine, Zhejiang, China,
[3]Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung 80424, Taiwan
himax26@zju.edu.cn,8013016@zju.edu.cn,himac26@zju.edu.cn,hwang@g-mail.nsysu.edu.tw

ABSTRACT

A local region model with attentive temporal-spatial pathways is proposed for automatically learning various target structures. The attentive spatial pathway highlights the salient region to generate bounding boxes and ignores irrelevant regions in an input image. The proposed attention mechanism allows efficient object localization, and the overall predictive performance is increased because there are fewer false positives for the object detection task for medical images with manual annotations. The experimental results show that proposed models consistently increase the base architecture's predictive performance on the Medico dataset with satisfactory computational efficiency.

1 INTRODUCTION

This study proposes a simple and effective solution that interfaces an attention mechanism in a standard CNN model. The feature maps are utilized more efficiently, and localization does not require processing the entire image. The proposed attentive model, which consists of tempo-spatial pathways, automatically learns to focus on target structures without additional supervision. The spatial pathway generates local region proposals on-the-fly using the salient features for a specific task. The temporal attention model proposes a sequence of locations for the local region search and not the entire image, so the computational overhead is significantly reduced, and many model parameters are omitted, similarly to multi-model frameworks. CNN models that use the proposed attentive model can be trained from scratch using standard methods or transfer learning. Similar attention mechanisms have been proposed for natural image classification and captioning [2, 4] for adaptive feature pooling, where model predictions are conditioned only using a subset of selected image regions. The proposed process assigns attention coefficients to specific local regions.

This study uses a novel hybrid attention model (HAM) as an interface between any feature extractors, such as a CNN, and a decision-making module for end-to-end tasks, such as RL, classification, regression. The proposed module determines spatial pinpoints in feature space using a hard attestation pathway. The model also synthesizes the context vector using a soft attention mechanism and a GRU for decision-making downstream. Real images are used to determine the efficacy of the proposed model and are used as a pre-training data set for detection and classification for colonoscopic images [6] that are the motif of this work. The contributions of this work are summarized as follows:

A hybrid attention approach allows an attention mechanism specific to local regions and the subsequent strategy or decision-making process. This improved model performs better than state-of-art methods that use global or local search schemes.

An attention interface is used for region proposals and sequential search of glimpses on local regions simultaneously for medical images. The proposed attention interface, which can be trained from end to end, replaces the hard-attention approaches currently used only for image classification. It eliminates the need for the global generation of bounding boxes for a Faster R-CNN [7] and provides better accuracy and greater computational efficiency than a local search scheme method. The study demonstrates that the proposed attention mechanism produces fine-scale attention maps that can be visualized with minimal computational overhead.

A masking scheme is applied to the distribution of attention scores to increase computational efficiency, instead of imposing directly on the feature map and influencing downstream operations. It ensures better classification performance than the baseline approach. It is shown that attention maps and an observation pinpoint allow fewer glimpses and fewer useful observations. A modification to the standard FPN is used for feature extraction, so the process is sensitive and specific.

2 APPROACH

2.1 Method

The process for the proposed local search method for polyps detection involves two stages [1]. During the first stage, the local region proposal network (RPN) proposes candidate ROIs from glimpsed regions located in sequence by the HAM. The weighted feature's attention scores are used to determine a glimpsed region in which target objects may reside. Bounding boxes are generated, and the process and the process then involves classification and position regression for preliminary screening. The confidence index for the classification is used to determine bounding boxes with higher values. Local non-maximum suppression is used to filter out some bounding boxes as regions of interest (ROIs), and these are used as inputs for the second stage network, which involves binding box regression and classification. When the RoIs are generated and accumulated in all the sequences for classification and bounding box

MediaEval'20, December 14-15 2020, Online

regression, an exhaustive search is initiated. This process involves considerable computing resources, so a method that uses a hybrid attention mechanism with RL to the RPN reduces calculation.

Instead of an exhaustive search over the entire image, the proposed method uses a Faster RCNN for a sequential search directed by a hybrid attention module (HAM) to determine glimpse regions that are likely to contain an object. RoI's are generated in a restricted area, where target objects are likely to be located. This local search reduces the amount of calculation for insignificant ROIs. The proposed model has four modules: a CNN-based feature extractor, the proposed HAM, a local RPN, and a detector for bounding box regression and object classification. Glimpse regions are pinpointed, and the length of the sequence of glimpses is determined sequentially. The local RPN generates bounding boxes of different sizes and aspect ratios within a glimpsed region. The detector regresses bounding boxes and classifies objects. The architecture of the HAM is shown in Figure 1.

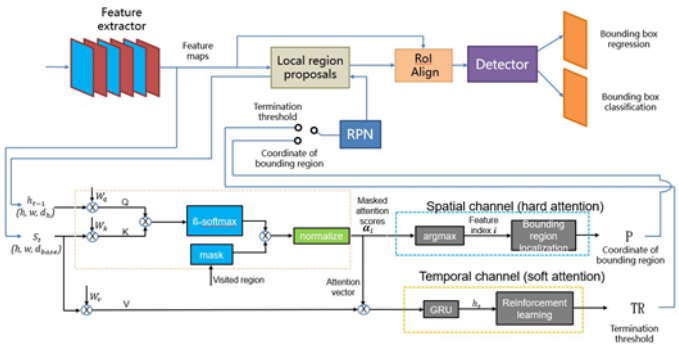

Figure 1: The architecture of the local region proposal method.

2.2 Preparation and Data set

The experiments were executed using the Ubuntu 18.04 operating system, Python 3.7, Tensorflow. The data sets for the experiments are provided in Medico Challenge [5]. A public data set of real scenes (PASCAL VOC [3]) is used to pre-train the Faster R-CNN framework. The data set contains only images, so data augmentation operations, such as rotation, reflection, and resizing, increase the number of images. Five-fold cross-validation is used for the experiments.

3 RESULTS OF COMPARISONS WITH PEER METHODS

The results for the colonoscopy dataset in Figure 2 show that the HAM-beta and HAM-beta-mask are similar to drl-RPN in terms of AP_50.There are fewer average glimpses and a smaller average glimpsed area than for the drl-RPN, and the AP density and glimpse contribution are better than peer methods.

The drl-RPN must search three times for important areas before terminating the glimpsing process, requiring more computation time. The HAM-beta and HAM-beta-mask accurately locate the correct in the first time search.

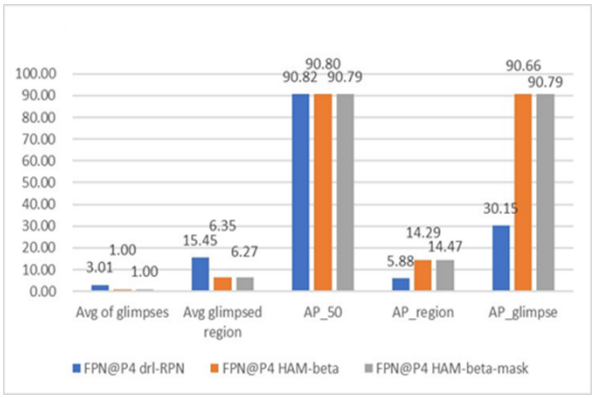

Figure 2: Comparisons between different configurations for the proposed model and peer methods.

4 CONCLUSION AND FUTURE WORK

This study proposes an innovative attention module that uses soft and hard attention. This module can interface with any architecture that involves simultaneous spatial and temporal tasks, such as polys detection. A global search scans the entire image in an object detection task, but it requires much time and resources. The proposed approach obviates the need to use an extra model by learning to highlight salient local regions in images. The proposed temporal-spatial attention module leverages the salient information in the state space for a policy learner, such as reinforcement learning, in addition to object detection in image tasks.

ACKNOWLEDGMENTS

This work is supported by the grant of the Key Project of Yiwu Science and Technology plan, China. No.20-3-067.

REFERENCES

[1] Sharib Ali, Felix Zhou, Barbara Braden, Adam Bailey, Suhui Yang, Guanju Cheng, Pengyi Zhang, Xiaoqiong Li, Maxime Kayser, Roger D Soberanis-Mukul, Shadi Albarqouni, Xiaokang Wang, Chunqing Wang, Seiryo Watanabe, Ilkay Oksuz, Qingtian Ning, Shufan Yang, Mohammad Azam Khan, Xiaohong W Gao, Stefano Realdon, Maxim Loshchenov, Julia A Schnabel, James E East, Georges Wagnieres, Victor B Loschenov, Enrico Grisan, Christian Daul, Walter Blondel, and Jens Rittscher. 2020. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports* 10, 1 (2020), 2748. https://doi.org/10.1038/s41598-020-59413-5

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88, 2 (2010), 303–338.

[4] Saumya Jetley, Nicholas A. Lord, Namhoon Lee, and Philip H. S. Torr. 2018. Learn To Pay Attention. *CoRR* abs/1804.02391 (2018).

[5] Debesh Jha, Steven A. Hicks, Krister Emanuelsen, Håvard Johansen, Dag Johansen, Thomas de Lange, Michael A. Riegler, and Pål

Halvorsen. 2020. Medico Multimedia Task at MediaEval 2020: Automatic Polyp Segmentation. In *Proc. of the MediaEval 2020 Workshop*.

[6] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. 2020. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*. Springer, 451–462.

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates, Inc.