ADULT IMAGE CLASSIFICATION BY A LOCAL-CONTEXT AWARE NETWORK

Xizi Wang, Feng Cheng, Shilin Wang, Huanrong Sun, Gongshen Liu, Cheng Zhou*

School of Electric Information and Electronic Engineering, Shanghai Jiaotong University

ABSTRACT

To build a healthy online environment, adult image recognition is a crucial and challenging task. Recent deep learning based methods have brought great advances to this task. However, the recognition accuracy and generalization ability need to be further improved. In this paper, a local-context aware network is proposed to improve the recognition accuracy and a corresponding curriculum learning strategy is proposed to guarantee a good generalization ability. The main idea is to integrate the global classification and the local sensitive region detection into one network and optimize them simulatenously. Such strategy helps the classification networks focus more on suspicious regions and thus provide better recognition performance. Two datasets containing over 150,000 images have been collected to evaluate the performance of the proposed approach. From the experiment results, it is observed that our approach can always achieve the best classification accuracy compared with several state-of-the-art approaches investigated.

Index Terms— adult image recognition, deep convolutional network, multi-tasks learning

1. INTRODUCTION

Early years of the 21st century have witnessed an explosion in the Internet usage. Internet offers people a universal access to much more information than ever before. Web images, as a major information transmission media, are increasing explosively. However, due to the lack of control over information sources, images with pornography or overexposure are widely spread online, which brings adverse effects to young people. Therefore, to build a healthy online environment, automatic detection of adult (pornography or sexy) images is a crucial and challenging task.

In the past decades, many researchers have proposed various approaches in adult image detection. Earlier works have focused on finding naked people by detecting skin regions and using color, texture and geometrical features [1]. Although this method is straightforward, limitations exist that not all images with large skin areas are related with pornography, e.g. face close-ups. Another popular kind of approaches are based on traditional image classification techniques [2]. With the development of deep learning, deep convolutional neural network (DCNN) based algorithms have also been adopted in this area [3][4]. The above methods can detect adult images with certain accuracy; however, various kinds of backgrounds, lighting conditions, human postures, etc. bring difficulties to the detection task. Moreover, in many image content rating systems, images with clothing overexposure or sexual intention (but not pornography) are also required to be inaccessible to children. This is a more challenging task [5] because some body photography, studio portrait shots, etc. are very similar to nude images.

In view of the above difficulties and challenges, a local-context aware classification network (LocoaNet) is proposed to classify images into three categories: benign, sexy and porn images. The main contributions of LocoaNet are three-folded. Firstly, both the global information from the entire image and the local information from the sensitive body parts are considered in the network and thus the LocoaNet can extract comprehensive and discriminative features to differentiate porn images, sexy images and benign images. Secondly, a multi-task learning scheme is designed, which can train the local object detection and global classification network simultaneously. Thirdly, a curriculum learning strategy [6][7] is proposed to facilitate retraining the LocoaNet on larger datasets.

2. RELATED WORKS

Automatic detection of adult images has been a major concern nowadays and is studied by many researchers. In general, the existing adult image detection approaches can be roughly divided into three categories as follows.

Skin-region-based approaches assume that adult images usually contain much more skin regions than benign images [8]. In 2011, Yin et al. [9] proposed a coarse degree texture filter and a fractal dimension-based geometry filter to detect skin pixels. Based on the skin regions detected, an adult image classification was performed accordingly. The major disadvantages of this kind of approaches are two-folded: i) In different scenes with various lighting conditions, the appearance of skin regions varies greatly, which will degrade the performance of skin detection; and ii) In some cases,

Xizi Wang, Feng Cheng, Shilin Wang*, Gongshen Liu are with School of Electric Information and Electronic Engineering, Shanghai Jiaotong University, 200240, Shanghai, China (e-mail: flxcc@qq.com, wscf1314@sjtu.edu.cn, wsl@sjtu.edu.cn). Huanrong Sun, Cheng Zhou are with SJTU-Shanghai Songheng Content Analysis Joint Lab. *Corresponding author

images containing many skin regions are benign images and vice versa, which may lead to inevitable classification errors.

Handcraft feature-based methods detect adult images based on various kinds of image features. In [5], the color features in terms of the color histogram and the color coherence vector are adopted as the discriminative features. Zhao and Cai [10] combined the color, edge and texture features with the SIFT features to enhance the detection performance. Compared with skin-region-based approaches, the feature-based methods usually achieve better detection accuracy and the detection performance is determined by the discriminative features employed.

Owing to the great achievements in image classification and understanding, many researchers have employed deep convolutional neural networks (DCNN) to detect adult images. Such methods are refered to as the **DCNN based approaches**. Moustafa et al. [11] combined AlexNet [12] with GoogleNet [13] as a model-ensemble. They refered to their network as AGNet and showed the detection accuracy of their model was slightly better than either single model. Ou et al. [14] considered the complementarity of global context and local context information and developed a context-ensemble detection system with fine-to-coarse strategy. Their approach is the first work to apply object detection in adult image recognition. They ensembled three contexts (object detection yields local and cross contexts, global classification yields global context) to make a final prediction. The ability to transfer to other datasets was limited because of their ensemble strategy, i.e. object detection network was trained alone and required sensitive body-part annotations.

3. THE PROPOSED METHOD

3.1. Overview of our method

Fig. 1 shows the overall architecture of the proposed LocoaNet. The network can be divided into three modules, i.e. the ResNet50 [15] backbone, the sensitive body part detection network (SpNet in short) and global classification network (GcNet in short). Note that the selection of ResNet50 is to balance between the classification accuracy and the processing time. Any other classical DCNN structure (e.g. VGG16 or ResNet101) can also be employed. The main factors that make LocoaNet outperform other methods are: i) the seamless integration of the local object detection and the global classification methods with multi-tasks learning; and ii) the curriculum learning strategy.

3.2. Sensitive body part detection network

The major drawback of the traditional global classification task in adult image detection is that the classification network tends to focus on global context such as background rather than discriminative regions such as naked bodies. For example, an image containing a naked woman in a small

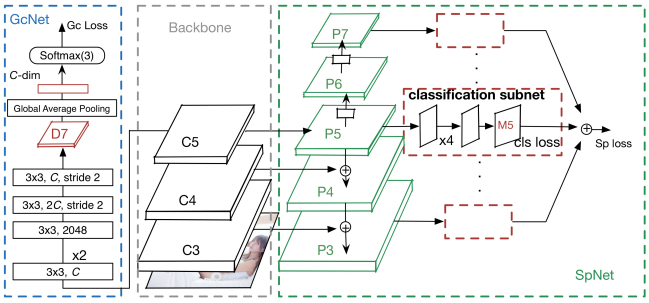

Fig. 1: The overall architecture of LocoaNet

region of a big prairie will probably be classified as a benign image. Therefore, a sensitive body part detection network (SpNet) is introduced to help the feature network focus more on sensitive body parts such as breast, ass and etc., which can better differentiate adult images.

Inspired by FPN [16] and RetinaNet [17], the proposed SpNet is designed to refine both the last feature layer and some intermediate layers. A feature pyramid is built on the top layers at each residual stage of the backbone network as shown in Fig.1. Following each level of the pyramid feature map P_i (i is ranged from 3 to 7), a number of convolution layers are applied to produce a feature map M_i with A channels. Each spatial location in M_i in a channel forms an anchor, which corresponds to an area in the input image centered at the location and of a specific shape determined by the channel parameter. All the anchors at each spatial location of M_i are analyzed to predict the presence of any sensitive body part class (in our approach, K=10 sensitive part classes are investigated). The anchor boxes settings are similar to those in FPN [16], i.e.: i) each anchor on the pyramid feature map M_i corresponds to an area of the size $2^i \times 2^i$; and ii) three aspect ratios, i.e. {1:2, 1:1, 2:1} and three scales $\{2^0, 2^{1/3}, 2^{2/3}\}$ are adopted (then A=3×3=9). The functionality of SpNet is to detect the anchors belonging to any sensitive body part class.

As shown in Fig.1, SpNet is composed of two parts, i.e. the pyramid feature maps, i.e. P_3 to P_7, and the classification subnets following each feature layer. For the feature maps, each P_i contains C channels (C is set to 256 in our experiments) and similar to that in [16], P_3 to P_5 are computed from top-down and lateral connections to the output of the convolution layers at each residual stage of the backbone network, i.e. C_3 to C_5 correspondingly. P_6 and P_7 are obtained via a convolution layer with a 3×3 kernel and a stride of 2. A ReLU activation layer is inserted between P_6 and P_7. For the classification part, the subnet is composed of four convolution layers (with a 3×3 kernel, same padding, C filters and ReLU-activation) for feature extraction and a convolution layer (with a 3×3 kernel, $K \times A$ filters and Sigmoid activation) for classification of A anchors at each spatial position into K sensitive part classes.

The main objective of incorporating the SpNet is to regularize the feature extraction layers related to C_3 to C_5. With

SpNet, the proposed LocoaNet can focus on suspicious regions and provide a more accurate classification result.

3.3. Global Classification Network

The global classification network classifies the images into three categories, i.e. {benign, sexy, porn}. GcNet is constructed on the last stage of backbone network. Taking C_5 as the input feature, a high level feature map D_7 (with the same width and height as those in P_7) is obtained via five convolutional feature extraction layers (whose parameters are given in Fig.1). All the convolution layers in GcNet employ the ReLU activation function. Then following a global average pooling layer, a three-unit full connection layer with softmax activation function is adopted for classification.

3.4. Training and Inference

Training: As stated in Subsection 3.2, SpNet aims to guide the feature layers in the backbone network to learn discriminative features. To achieve this goal, a multi-tasks learning scheme is proposed. The overall loss is the sum of two parts, i.e. the detection loss and the classification loss. The focal-loss [17] is adopted to calculate the detection loss. In the training stage, anchors are assigned to ground-truth object boxes if their intersection-over-union (IoU) is greater than 0.5 and to background if IoU is in the range of [0,0.4). Anchors whose IoU is in the range of [0.4,0.5) are ignored during training. On the other hand, the cross-entropy is adopted to compute the classification loss.

The backbone network is initialized with pre-trained weights on ImageNet. SpNet is initialized using the same approach as RetinaNet [17]. For layers in GcNet, the kernels are initialized with glorot_uniform [18] and bias with zeros. LocoaNet is trained with Adam optimizer [19] with a learning rate of 10^{-5}, a clipnorm of 10^{-3} and other default parameters as in [19]. The input image is resized to min-side of 600 and max-side of 1024 while keeping the aspect ratio.

Inference: Given an image to inference, SpNet is inactivated and only the backbone network and GcNet are functional, which makes the computational complexity of our approach comparable to that of the backbone network in the inference stage.

3.5. Curriculum learning strategy

It is widely known that when predicting images outside the training dataset, there will be a performance degradation for most deep learning paradigms [20, 21]. The transfer learning ability to a larger dataset is of great importance.

In our application, manually annotating the ground truth of the sensitive body part regions in the image is much more time consuming than labelling the image class, which will reduce the scalability of the proposed approach. To solve this problem, a curriculum learning strategy is proposed to extend our model to a larger dataset without any additional region-annotated training sample, which runs as follows.

Step 1: Train LocoaNet using the original sensitive-part annotated dataset (referred to as the original set hereafter). Two networks, i.e. GcNet and SpdNet are initialized and share the same ResNet50 backbone.

Step 2: Based on the extension training dataset where only the class label is available (referred to as the extension set hereafter), ignore SpNet and update the weights in the backbone network and GcNet following the forward and backward optimization strategy, aiming to minimize the classification loss.

Step 3: Based on the original set, ignore GcNet and update the weights in both the backbone network and SpNet, aiming to minimize the detection loss.

Step 4: Repeat Step 2 & 3 until converge.

Batch-learning is utilized in step 2&3. The optimizers of both networks are SGD with a learning rate of 10^{-5} , a weight decay of 10^{-8} and a momentum of 0.9. The underlying idea of our strategy is that during optimization, limit the search space to a much smaller solution space restricted by SpNet.

4. EXPERIMENTS AND DISCUSSIONS

To comprehensively evaluate the performance of the proposed method, two datasets, i.e. our AIC dataset and the public NPDI [22] are adopted in the experiments and the detailed descriptions of the datasets are given as follows.

AIC dataset: Since there is no public available dataset for adult image content rating, we built a three-classes (i.e. porn, sexy and benign images) dataset, which is referred to as the Adult Image Classification (AIC in short) dataset. The definitions of each category are similar to [14]. Some sample images of each class are given in Fig.2. There are 150,000 images (benign 50,000, sexy 50,000 and porn 50,000) with correct class labels in the dataset and 80% benign images contain one or more people, which makes the benign images more similar to the sexy/porn ones. Moreover, 14,000 images containing manual annotated sensitive regions and 8,000 benign images in the AIC dataset are selected to construct the AIC_R subset. Ten sensitive-part classes are annotated in AIC_R, specifically, {breast_porn, vulva_porn, dick_porn, pubes_porn, ass_porn, breast_sexy, ass_sexy, back_sexy, frontleg_sexy, body_sexy}.

NPDI: The pornography dataset contains 400 porn and 400 benign videos and it is adopted to evaluate the two-class classification performance. For the benign class, 200 videos are randomly chosen (called “easy”) and 200 videos are

Fig. 2: AIC samples

Table 1: evaluation on AIC_R dataset

	Precision(%)			Recall(%)			Acc(%)
	B	S	P	B	S	P	
ResNet50	91.3	92.5	94.3	97.1	88.4	92.4	92.6
GcNet	92.0	93.3	95.2	97.6	89.4	93.6	93.5
LocoaNet	**95.1**	**98.8**	**96.7**	**99.6**	**90.4**	**99.0**	**96.3**

Table 2: evaluation on AIC dataset

	Precision(%)			Recall(%)			Acc(%)
	B	S	P	B	S	P	
GcNet	92.1	91.5	**98.0**	94.7	91.1	95.6	93.8
LocoaNet (finetune)	92.9	91.1	97.7	94.4	91.8	95.5	93.9
LocoaNet(curriculum)	**93.9**	**96.6**	96.9	**97.4**	**92.2**	**97.7**	**95.8**

Table 3: Methods are trained on AIC.

Methods	AIC							NPDI					
	Precision (%)			Recall (%)			Acc (%)	Precision(%)		Recall(%)		Acc (%)	Time (ms)
	B	S	P	B	S	P		B	P	B	P		
ResNet50	**94.1**	92.5	94.4	94.0	91.2	95.1	93.4	91.3	83.8	89.8	86.1	88.4	38
AGNet	90.9	89.5	91.2	92.3	87.3	92.1	90.5	87.2	83.6	90.6	78.2	85.9	**21**
DMCNet	92.7	92.4	93.2	93.6	90.7	94.1	92.8	90.7	87.6	92.7	84.5	89.6	119
LocoaNet (curriculum)	93.9	**96.6**	**96.9**	**97.4**	**92.2**	**97.7**	**95.8**	**92.5**	**91.6**	**95.1**	**87.4**	**92.2**	40

selected from textual search queries like "beach", "wrestling" (called "difficult"). From the videos, 16,727 key frames are extracted for evaluation.

The training/validation/test sets contain 80%/10%/10% samples for evaluations on the AIC and AIC_R datasets. In NPDI, 100 non-porn and 40 porn videos (1591 frames) are randomly chosen to adjust model parameters. Other images in NPDI are adopted as test samples and images classified as sexy/benign by our model are regarded as benign samples. Ten random tests are performed for each evaluation and the average result is recorded to avoid performance variations caused by different selections of the training set.

4.1. Integration of SpNet

To evaluate the effectiveness of the integration of the local SpNet and the global GcNet, the following experiments have been carried out on our AIC_R dataset. Three networks have been investigated, including the ResNet50 as the baseline, our GcNet alone and our LocoaNet, and the results are given in Table 1 (where B, S and P denote the benign, sexy and porn class, respectively). From the table, it is observed that i) the proposed GcNet alone performs slightly better than the baseline and ii) with the integration of the SpNet, an accuracy gain of 2.8% have been achieved, which demonstrates the effectiveness of our network structure.

4.2. Evaluation of the Transfer Learning Ability

The ability to transfer to large dataset is of great importance in many machine learning applications. In our application, the transfer learning ability is evaluated by extending the model from the region annotated AIC_R to the entire AIC dataset. The following three approaches have been evaluated and the results are listed in Table 2: 1) Retrain the GcNet on AIC as the baseline; 2) Finetune LocoaNet on AIC with pretrained weights on AIC_R; and 3) Extend our model by the curriculum learning strategy.

From Table 2, it is observed that the GcNet retraining and the LocoaNet finetuning approaches achieve comparable results and with the curriculum learning strategy, an accuracy gain of 1.9% is achieved. Compared with that on the small AIC_R dataset (listed in Table 1), the classification performance on the AIC dataset does not degrade much, which demonstrated the effectiveness of the proposed curriculum learning strategy.

4.3. Comparison with state-of-the-art methods

To comprehensively evaluate the performance of the proposed method, two state-of-the-art methods, i.e. the AGNet [11], DMCNet [14] and the baseline ResNet50 [15] with large input size (min-side 600 and max-side 1024) are adopted for investigation and the experiment results are given in Table 3.

From the table, the following issues can be concluded: i) for all the three classes, the sexy class is more confusing for all the approaches investigated than the benign and porn classes (with lower precision/recall). It is mainly because some sexy images and porn images are quite similar and in some cases, it is difficult for even human beings to determine the correct class label; ii) An accuracy gain of 2.6% on NPDI against the second best DMCNet which also considers the local information. It can be contributed to the seamlessly incorporation of the SpNet and the multi-task learning scheme, which helps LocoaNet extract features with highly discriminative power; iii) Among all the approaches investigated, the proposed LocoaNet achieves the best performance on both datasets.

5. CONCLUSION

In this paper, a local-context aware classification network (LocoaNet) is proposed for adult image classification. By incorporating a local sensitive body part detection network and employing a multi-tasks learning scheme, our LocoaNet can extract highly discriminative features for adult image classification. A curriculum learning strategy is also proposed to improve the generalization ablility. Experiment results have demonstrated the proposed approach can achieve satisfactory detection accuracies with a high processing speed in both datasets investigated.

6. ACKNOWLEDGMENT

The work described in this paper is supported by NSFC Fund (No. 61771310) and program of Shanghai Technology Research Leader under grant 16XD1424400.

7. REFERENCES

[1] Feng Jiao, Wen Gao, Lijuan Duan and Guoqin Cui, "Detecting adult image using multiple features," *2001 International Conferences on Info-Tech and Info-Net. Proceedings (Cat. No.01EX479)*, Beijing, pp. 378-383 vol.3, 2001.

[2] M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection," *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, Fort Collins, CO, pp. 280 Vol. 1, 1999.

[3] Nian, Fudong, Teng Li, Yan Wang, Mingliang Xu, and Jun Wu. "Pornographic image detection utilizing deep convolutional neural networks." *Neurocomputing 210*, pp.283-293, 2016.

[4] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.

[5] D. Ganguly, M. H. Mofrad and A. Kovashka, "Detecting Sexually Provocative Images," *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, pp. 660-668, 2017.

[6] Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston. "Curriculum learning." *In Proceedings of the 26th annual international conference on machine learning*, pp. 41-48. ACM, 2009.

[7] Sarafianos, N., Giannakopoulos, T., Nikou, C., & Kakadiaris, I. A. "Curriculum Learning of Visual Attribute Clusters for Multi-Task Classification". *arXiv preprint arXiv:1709.06664*, 2017.

[8] Ries, Christian X., and Rainer Lienhart. "A survey on visual adult image recognition." *Multimedia tools and applications 69, no. 3*, pp.661-688, 2014.

[9] H. Yin, X. Xu and L. Ye, "Big Skin Regions Detection for Adult Image Identification," *2011 Workshop on Digital Media and Digital Content Management*, Hangzhou, pp. 242-247, 2011.

[10] Z. Zhao and A. Cai, "Combining multiple SVM classifiers for adult image recognition," *2010 2nd IEEE InternationalConference on Network Infrastructure and Digital Content*, Beijing, pp. 149-153, 2010.

[11] Moustafa, Mohamed. "Applying deep learning to classify pornographic images and videos." *arXiv preprint arXiv:1511.08899*, 2015.

[12] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In Advances in neural information processing systems, pp. 1097-1105, 2012.

[13] C. Szegedy *et al*., "Going deeper with convolutions," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp. 1-9, 2015.

[14] Ou, Xinyu, Hefei Ling, Han Yu, Ping Li, Fuhao Zou, and Si Liu. "Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-Coarse Strategy." *ACM Transactions on Intelligent Systems and Technology (TIST) 8, no. 5*, pp.68. 2017.

[15] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, pp. 770-778, 2016.

[16] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, pp. 936-944, 2017.

[17] *T. Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal Loss for Dense Object Detection," 2017 IEEE International Conference on Computer Vision (ICCV),* Venice, pp. 2999-3007, 2017.

[18] Glorot, Xavier, and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks." *In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249-256. 2010.

[19] Kingma, Diederik, and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980*, 2014.

[20] Khosla, Aditya, Tinghui Zhou, Tomasz Malisiewicz, Alexei A. Efros, and Antonio Torralba. "Undoing the damage of dataset bias.", *European Conference on Computer Vision, Springer, Berlin*, *Heidelberg*, pp. 158-171, 2012.

[21] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," *CVPR 2011*, *Providence, RI*, pp. 1521-1528, 2011.

[22] Avila, Sandra, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo De A. AraúJo. "Pooling in image representation: The visual codeword point of view." *Computer Vision and Image Understanding 117, no. 5*, pp.453-465, 2013.