参赛队员姓名：胡雨森

中学：北京德威英国国际学校

省份：北京

国家/地区：中国

指导教师姓名：Yosef Karasik

论文题目：Solving Pediatric Vehicular Heatstroke with Efficient Multi-Cascaded Convolutional Networks

本参赛团队声明所提交的论文是在指导老师指导下进行的研究工作和取得的研究成果。尽本团队所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果。若有不实之处，本人愿意承担一切相关责任。

参赛队员：胡雨森　　指导老师：Yosef Karasik

2020 年 09 月 13 日

Solving Pediatric Vehicular Heatstroke with Efficient Multi-Cascaded Convolutional Networks

Yusen Hu

August 2020

Abstract

Pediatric Vehicular Heatstroke (PVH) is the situation where children suffer fatal injuries due to heatstroke after being forgotten in vehicles. It is a severe social problem: According to incomplete statistics, at least 864 children have died due to PVH since 1998 in the USA alone, and another 22 have lost their lives in 2020. In this paper, we developed a machine-learning based embedded warning system that mitigates such tragedies. Specifically, we present our Children in Vehicles (CIV) dataset, where we collected 2,076 positive samples of children and 1,529 negative samples of empty car interiors. We then present the framework and training process of our multi-cascaded convolutional network architecture that can detect children with a 98% accuracy. Furthermore, we demonstrate the power of our novel curriculum learning method, which improved the classification accuracy of our facial age estimator from 46% to 62% and its F1 score from 0.66 to 0.91. We also deployed our complete pipeline onto an embedded platform to present its overall feasibility. Additionally, we open-sourced our code and dataset for others to use & experiment with.

Keywords: multi-task learning, cascaded networks, convolutional neural networks, curriculum learning, Pediatric Vehicular Heatstroke

Contents

1 Introduction

Pediatric Vehicular Heatstroke (PVH) [1] is a severe social problem where children suffer fatal injuries due to heatstroke after being forgotten in vehicles. According to incomplete statistics, at least 864 children have died due to PVH since 1998 (shown in Figure 1), and another 22 have lost their lives in 2020 [1]. Various studies have demonstrated a large portion of PVH cases are not due to negligence, but a neurobiological condition known as the "forgotten baby syndrome" [2]. Although many organizations [1], [3], [4] have been constantly promoting public awareness of this issue, the number of severe incidents of PVH has not decreased; on the contrary, there is a slow upward trend.

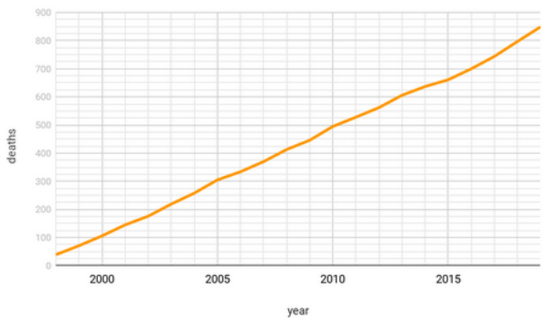

Figure 1: Cumulative Pediatric Vehicular Heatstroke deaths since 1998 in the USA [1]

We aim to develop a deep neural network model for the embedded device that is capable of detecting the presence of children in vehicles, relying solely on visual information. However, training such a model is a challenging task: *1)* There are no existing datasets suitable for our research purpose. Most existing datasets have insufficient images of children and many lacked variation in face direction, illumination, and occlusion. *2)* Since our model uses visual information to detect the presence of children, it faces the same challenges that computer vision algorithms face in general, such as illumination, occlusion, deformation, background cutter, intraclass variation, and so on. *3)* Our model needs to determine whether a detected human is a child or an adult, since adults are irrelevant to the PVH problem. The model's inability to determine the differences would lead to unnecessary false alarms. *4)* It is difficult to deploy machine learning models onto embedded devices. To the best of our knowledge, state of the art models are often too large and cumbersome for embedded platforms due to limitations in memory and computing power.

To solve the aforementioned problems, we first constructed our Children in Vehicles dataset, which has 2,076 samples of children in cars and 1,529 negative samples. We then designed and trained a lightweight, multi-task cascaded convolutional network that detects the faces/bodies of children in an image, and used curriculum learning to train our age classification network. Additionally, we deployed our model onto an embedded device to test its usability.

In summary, the major contributions of this paper are listed as follows:

1. We built the Children in Vehicles (CIV) dataset, which, to the best of our knowledge, is the first dataset built for addressing the PVH problem.

2. We designed an efficient, multi-task cascaded convolutional network architecture, which occupies less than 4.7MB of space, for detecting children in cars and classifying the age of detected faces to lower false-alarms.

3. We verified the effectiveness of our multi-task approach to simultaneously detect the faces and bodies of children, which improves the model's ability to apply learned patterns to unseen data.

4. We demonstrated the usefulness of our novel curriculum learning method, which improved the overall accuracy of our age classification model by 16%.

5. We deployed our pipeline onto an embedded platform to evaluate its overall feasibility.

The rest of the paper is organized as follows: other work related to our research is first introduced in Section 2. Our dataset collection & processing details is outlined in Section 3. We then present our framework, model architecture, and implementation details in Section 4. Finally, training procedures and experiment results are discussed in Section 5.

2 Related Work

Our research is based on convolutional neural networks, multi-task learning, and face detection algorithms.

Convolutional Neural Networks (CNN): After the success of AlexNet [5] by Krizhevsky et al., CNNs have gained widespread popularity in solving various computer vision tasks. Many innovative CNN architectures have been developed subsequently, such as VGG [6], ResNet [7], or Inception [8]. The trend in these works has been to make deeper networks to achieve better performance. However, these advancements are not necessarily making networks more efficient with respect to size and speed. To address such issues, architectures such as DenseNet [9] are created, where inter-layer and skip connections are increased and feature reuse is maximized, instead of simply building a deeper and wider network for improving performance. These techniques are not used in our research, due to their large size and high computational costs.

With the rise in demand for deep learning on embedded devices, lightweight models such as MobileNets [10] are created, which are designed specifically for embedded vision applications. It uses depthwise-separable convolutions to replace regular convolution operations, which leads to a smaller model (due to the fewer number of parameters) and a lower complexity (due to the fewer number of multiplications) while maintaining comparable performance. As such, mobilenets are commonly deployed onto edge devices, where computational resources are relatively constrained. However, the depthwise-separable convolution technique in MobileNets is unsuitable for our task. Its size-reduction effects are not substantial for small networks, and may even hurt the model's performance by decreasing the number of learnable parameters.

Object Detection: Many state-of-the-art object detection models also stemmed from the development of CNNs. Two-stage detectors (R-CNN [11], Fast R-CNN [12], Faster R-CNN [13], Mask R-CNN [14], etc.) or one-stage detectors such as YOLO [15] all use some form of CNN as the object classifier. R-CNN and its variants are too large for deployment onto an embedded device, while YOLO makes sacrifices in accuracy for real-time performance, which is the opposite of our goal.

Face Detection: In the early eras of computer vision, face detection methods usually use traditional image processing, such as edge detection, template matching, and so on [16]. However, these were soon outclassed by deep learning-based approaches. The Multitask Cascaded Convolutional Networks (MTCNN) [16] is a popular example. Through multitask learning, MTCNN unifies the two tasks of facial detection and alignment into one to exploit

the inherent connections between them, which improves its performance. The network consists of three stages (P-Net, R-Net, and O-Net) that detects & aligns faces in a coarse-to-fine manner, with non-maximum suppression between the stages to eliminate redundant bounding boxes. We use the PNet and ONet architecture from MTCNN as they are shown to be more computationally efficient and achieve better performance than similar networks.

Age Estimation: Stemming from advances in CNNs and face detection, many deep learning-based age estimation models are proposed, such as DEX [17], CORAL [18], or DLDL [19]. However, all these models use state-of-the-art CNN architectures such as VGG and are concerned with the task of estimating age in the continuous range from 0 to 100, which makes them unfit for our task of classifying children and adults on an embedded system.

3 Children in Vehicles (CIV) Dataset

3.1 Data Collection

To the best of our knowledge, there are no existing datasets that are suitable for our research purpose. Most existing datasets have insufficient images of children or images in a vehicular setting, and many lacked variation in face direction, pose, illumination, and occlusion. Thus, we are motivated to build our own Children in Vehicles (CIV) dataset for training and evaluating our model that detects children in vehicles. We collected images of children below the age of 7, because children over 7 years are less than 2% of PVH cases in the USA.

With the help of 68 volunteers, we collected 3,605 images for our CIV dataset. We guided them on how to take photos suitable for the dataset, which includes instructions on how to vary the photography angle, distance, the child's pose, environment lighting, and so on. We also asked the volunteers to collect negative samples of vehicle interiors where children aren't present, since these are needed for training our model. Due to time and other constraints, we did not collect negative samples where adults are present in the vehicle. In the end, we obtained 2,076 positive samples and 1,529 negative samples. Approximately 20% of the positive samples are children whose frontal face is not visible. All of the children in the dataset are Asian.

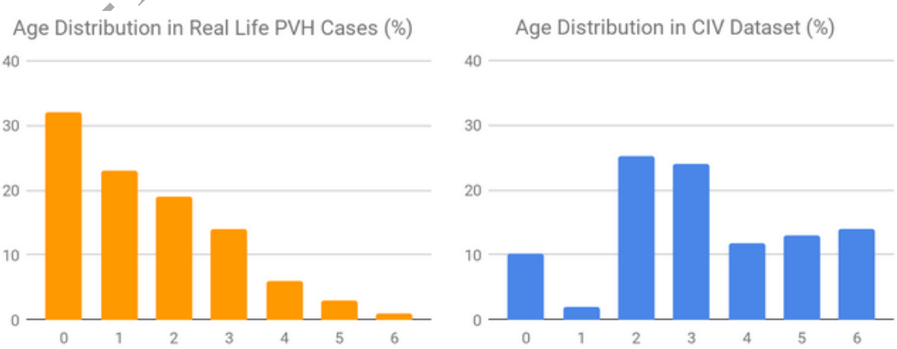

Figure 2: Age distribution in PVH cases in the USA [1], and in our CIV dataset

We tried our best to keep the number of samples in different age categories evenly distributed to ensure our model learns to recognize children of all ages, but unfortunately there were few volunteers at the time who had children below the age of 2. The age distribution histogram of our dataset is shown in Figure 2.

As shown in Figure 3, our CIV dataset reflects real-world scenarios under which PVH occurs. Children are photographed with a variety of poses (many of which are natural) and

Figure 3: Sample images from our dataset. Row 1 to 4 shows the CIV dataset's pose & photography angle variation, wide range of occlusion, background variety, and diverse lighting conditions respectively.

from different angles, which simulates the possible locations that an embedded warning device could be placed. In real-life situations, children often sleep on the cars and are commonly covered to some degree by safety harnesses or comforters, which is reflected by our dataset's wide range of occlusion in the images. The dataset also contains images taken under a variety of illumination conditions, such as night time, backlit, over-exposure, partial illumination, and shadows. These correspond to less-than-ideal situations in vehicles, e.g. during noon, when the sun shines directly into the car. Through reflecting real-life scenarios, our CIV dataset enables models to generalize learned features and provides an accurate metric of a model's performance.

3.2 Data Processing

The raw images were renamed according to the children's age and resized to 400 by 400 pixels (some images had minor aspect ratio distortions). The LabelImg [20] python tool was used to manually annotate the bounding boxes for the child's body in each image. The bounding box includes the child's head, arms, and torso. The legs were excluded because they aren't visible in many images and in real-life scenarios, they are easily occluded.

For labeling faces, we used a pretrained MTCNN model to automatically generate annotations. We manually checked the results to verify MTCNN's automatic annotation quality. Several problems were noticed: firstly, there were some instances where MTCNN falsely labeled a negative region containing limbs or flesh-like colors as a face.

In order to fix this, we modified MTCNN's default detection threshold from 0.6, 0.7, 0.7 to 0.7, 0.8, 0.8, which allows MTCNN to filter out regions with a low confidence score and eliminate as many false positives as possible (we removed the remaining few false positives through another round of manual inspection). However, this introduced a new issue: there are now instances where obvious frontal faces are ignored. Out of the 408 cases where MTCNN failed to detect a face, 5 of them were false negatives. There were also 46 images

where negative regions were falsely labeled as faces in addition to correctly-labeled faces, creating more than one annotation per image. These are removed through manual inspection of the dataset. In the end, we stored our 2,076 annotations for the positive samples into binary NumPy array formats, for saving storage space and the ease of parsing/editing.

The CIV dataset is available for download from Google Drive: `https://drive.google.com/file/d/1R-JBM8pj8V9oC1AKwbGo5T7fRH4fTMoN/view?usp=sharing`.

4 Our Method

4.1 Framework

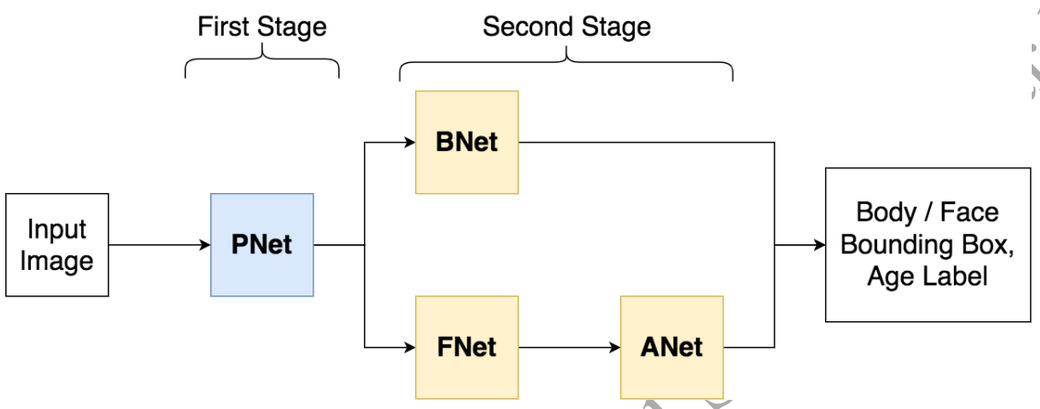

Figure 4: Our two-stage detection pipeline

As shown in Figure 4, our detection pipeline consists of four sub-networks, PNet, BNet, FNet, and ANet, arranged in a two-stage fashion. The first stage generates coarse region proposals for bodies and faces, while the second stage performs more precise classifications.

For an input image, we first create an image pyramid through repeatedly downsizing the image by a factor of 0.709 until the image is smaller than 12x12 pixels. The purpose of creating an image pyramid is to allow our network to detect features of varying sizes (heavily downsized images allows the network to detect children who are closer to the camera, while images that have higher resolution allows the network to detect children that are further away from the camera).

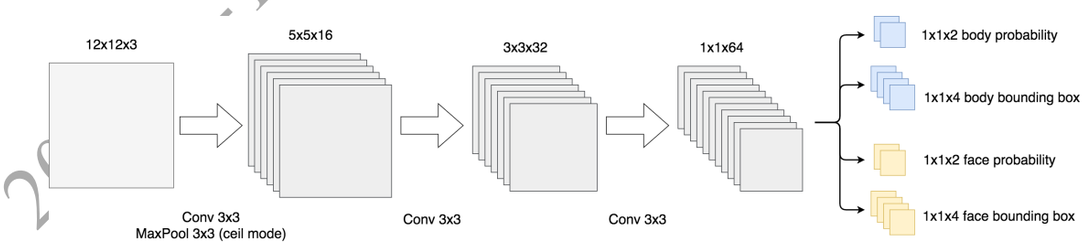

Figure 5: Our PNet architecture. Convolution and maxpooling operations have a stride of 1 and 2 respectively.

We based the first stage (shown in Figure 5 of our cascaded model on the MTCNN PNet [16] architecture, which generates "proposals" of regions that possibly contain someone's face or body. Our PNet is modified into completing four tasks simultaneously, namely face probability & bounding box regression, and body probability & bounding box regression. PNet is a fully-convolutional network, meaning it can accept input of any size above 12x12. It essentially replicates a "kernel" of size 12x12 that moves with a stride of 2 across an input image. Each 1x1 area in the output of PNet contains the extracted information of any potential faces/bodies present in the corresponding 12x12 area in the input image. Since

any faces/bodies are significantly larger than 2 pixels, a stride of 2 will have no accuracy impacts while reducing computational costs.

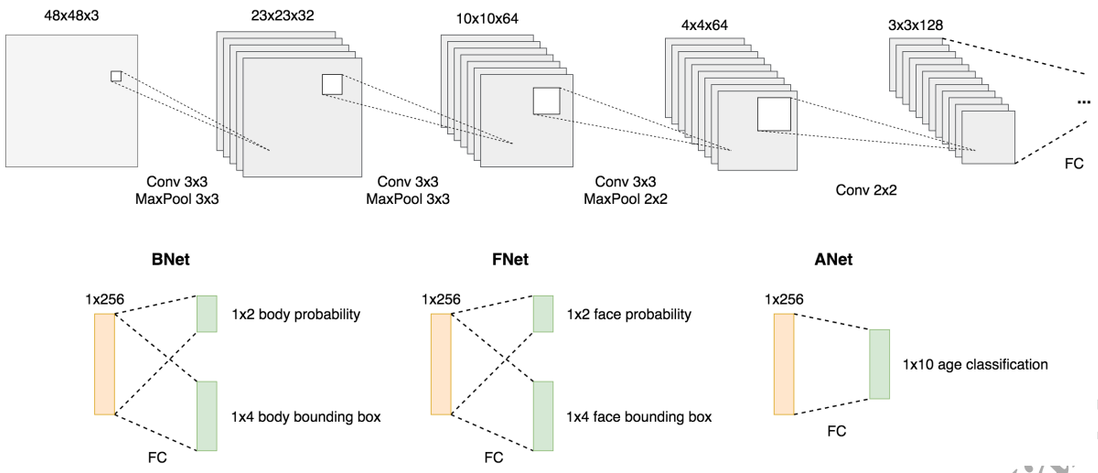

Figure 6: Our BNet, FNet, and ANet structure. They use the same convolutional layers but different task-specific fully-connected layers at the end.

The second stage of our cascaded model (shown in Figure 6) no longer uses the multitask-learning framework. Instead, we split the second stage into three separate models (BNet, FNet, and ANet), all of which are copies of the proven and robust MTCNN ONet [16] architecture (except for the FC layer of ANet). The structure of our 2nd-stage networks is shown above. The four networks within our model are trained separately, then combined together to form our final detection pipeline.

4.2 Implementation Details

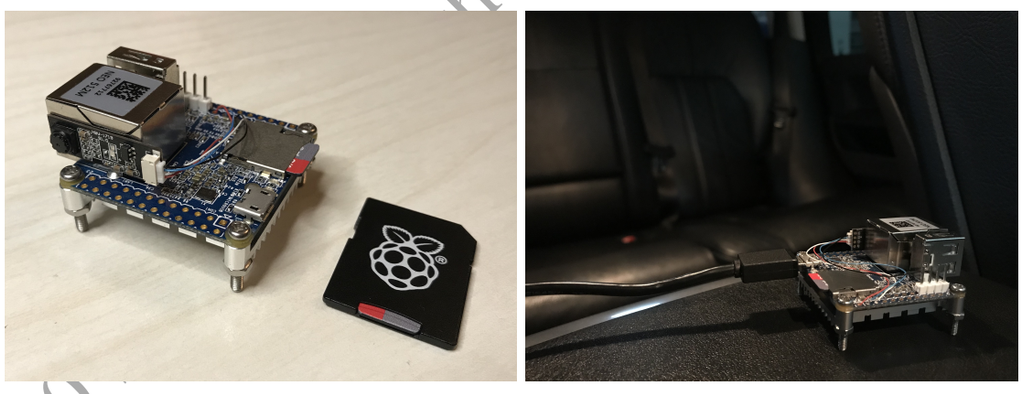

Figure 7: Left: Embedded device prototype (SD card for size reference). The SMS module (not visible) is located under the board. Right: The device is positioned on the center console of the car, facing the backseat. It is powered through the car's auxiliary power outlet.

Our entire detection pipeline is implemented in unoptimized Python3 code, utilizing external libraries including PyTorch [21], NumPy [22], and so on. The code to our project can be downloaded from GitHub: `https://github.com/Unturned3/mccn`. Our prototype embedded warning device is composed of several hardware modules: an image capture device (standard pinhole camera), a single-board computer[1], and a generic SMS module for sending text messages. Similar to how emergency lights work, the device will sleep & charge when the car is supplying power and activate if the power is cut (due to the car engine being switched off). Images taken by the camera will be processed by our pipeline, and if children are

[1] `http://wiki.friendlyarm.com/wiki/index.php/NanoPi_NEO`

deemed to be present, the vehicle's owner will receive text message warnings. Our prototype device is shown in Figure 7.

We executed our detection pipeline on the embedded platform and measured its inference performance. For each input image, it takes on average 114 millisecond to generate a result, which is equivalent to roughly 8.7 frames per second. In the case of addressing PVH, we do not need a real-time model.

5 Training & Experiments

5.1 Data Augmentation

PNet was trained directly on the CIV dataset, with data augmentation techniques employed. We randomly shuffled the images in the CIV dataset and split them into 80% and 20% for the training and validation set respectively. For each original image in the training set, we generate additional samples according to the following:

1. Generate additional negative samples: A square of side length x is repeatedly used to randomly crop the original image, until this cropped area has an IOU less than 0.1 with the labeled body and less than 0.05 with the labeled face. We deem such an area as a negative sample and append it to the training set. We then repeat the same steps, but decrease x by a factor of 0.709. We begin the generation process with $x = 100$ and stop when $x < 12$.

2. Generate additional positive body samples: This is only applicable to images that contain children. A square of side length x is used to randomly crop the original image a maximum of 20 times. In each iteration, we calculate the cropped area's IOU with the body labeled in the original image:

   (a) $0.0 \leq \text{IOU} < 0.1$: not a valid positive body sample. Repeat.

   (b) $0.1 \leq \text{IOU} < 0.2$: 20% body confidence

   (c) $0.2 \leq \text{IOU} < 0.4$: 50% body confidence

   (d) $0.4 \leq \text{IOU} < 0.7$: 80% body confidence

   (e) $0.7 \leq \text{IOU} < 1.0$: 100% body confidence

   In the case of b, c, d, e, we add the cropped image to the training set as a positive body sample and stop. A maximum of 20 tries is enforced because for some images with partial or small children bodies, it is impossible to get an IOU > 0.1. We initialize x as 300. Each side length x is used to generate three samples. x is then scaled down by a factor of 0.95. We stop when $x < 200$.

3. Generate additional positive face samples: This only applies to images that contain labeled faces. We start with a square of size 250 and stop when it's smaller than 150, scaling by 0.9 every time. The square's random location is confined in such a way that it always fully contains the originally labeled face's bounding box. Since the body's bounding box always contains the face's bounding box, each additional positive face sample we generate here will also contain a label for the body's bounding box, cropped appropriately.

Such augmentation techniques were proven to be effective, when we integrated PNet into the full detection pipeline by running it over an image pyramid. Figure 8 exemplifies the

difference between PNet trained without data augmentation (left), and PNet trained with augmentation (right). It is clear that the false positive rates were significantly reduced.

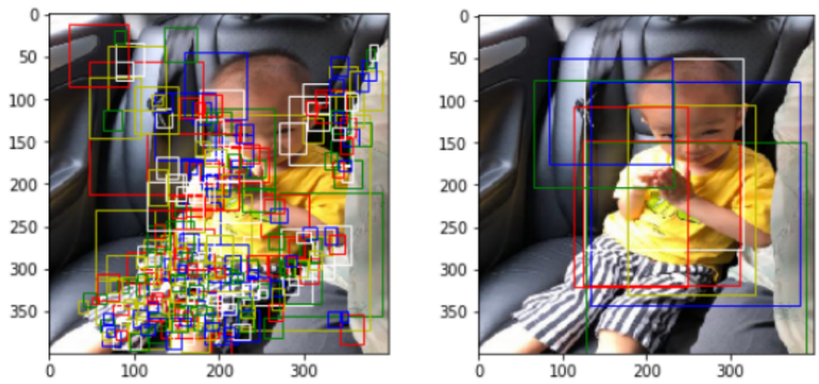

Figure 8: PNet performance without and with data augmentation. The bounding boxes' colors are varied for the purpose of visual clarity.

Additionally, we eliminated the first few high-resolution images in the image pyramid and only kept the last 6 (the most-downsized ones), which further reduced the number of false positives generated as well as computational costs of the overall pipeline. In a vehicular environment, the camera (on the embedded device) can only be placed at a certain distance away from the backseat. From the images we collected, we observed that children are guaranteed to occupy at least of the image's size. Thus, there is no need for the high-resolution images in the pyramid since there are no cases where the children could be that small.

5.2 Multi-Task Learning

To demonstrate the effectiveness of our multi-task framework, we trained three versions of PNet:

1. Body detection only, which ignores the loss function of the face detection tasks
2. Face detection only, which ignores the loss function of the body detection tasks
3. Multi-task, which trains both the body and face detection tasks

It is observed that PNet in both 1 and 2 generally converged to a better training loss than 3 in 50 epochs, as shown in Fig. 9. However, the multi-task version performed better on the validation dataset, as shown in Figure 9.

Table 1: Validation precision and recall for 3 PNet versions. The best value in each column is written in bold.

PNet Type	Body Detection Precision	Body Detection Recall	Face Detection Precision	Face Detection Recall
Body only	0.98	0.96	0.69	0.12
Face only	0.68	0.91	**0.92**	0.91
Multi-task	**0.98**	**0.98**	0.87	**0.95**

It has been shown that multi-task learning can act as a regularizer that prevents the model from overfitting [23], [24]. This is clearly demonstrated in our results: non-multitask PNets achieved better training losses but performed worse on the validation set.

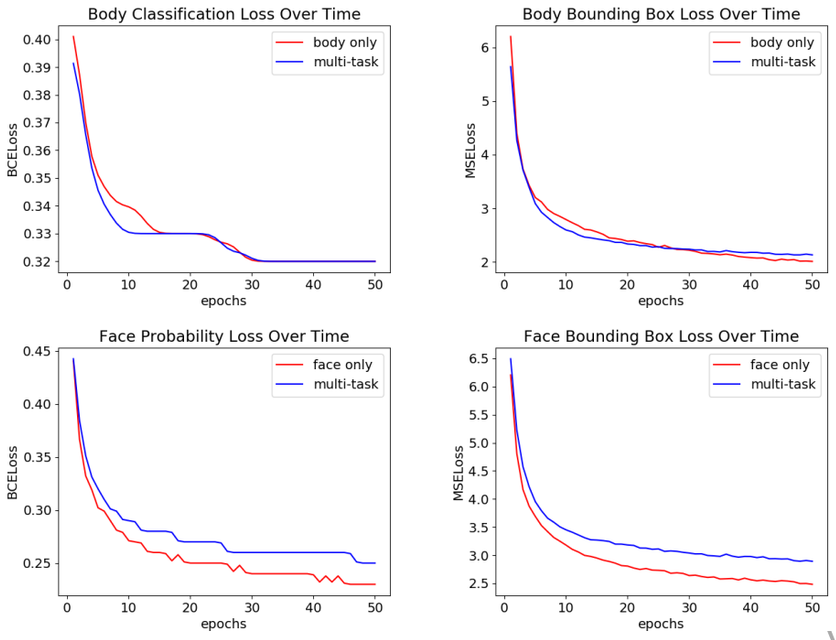

Figure 9: Training loss over time for the 3 PNet versions

We trained BNet in a similar fashion to compared to the PNet training procedure, except we disabled the generation for extra face samples and altered the program to generate more magnified crops additional children body samples, since BNet will receive the region proposals from PNet, which are usually close-ups of children. BNet is trained for 120 epochs and achieved a 98% accuracy in identifying children bodies.

The FNet in our network directly uses the parameters from a pretrained MTCNN ONet. Our CIV dataset has limited face samples and would be insufficient for training a face detector from scratch.

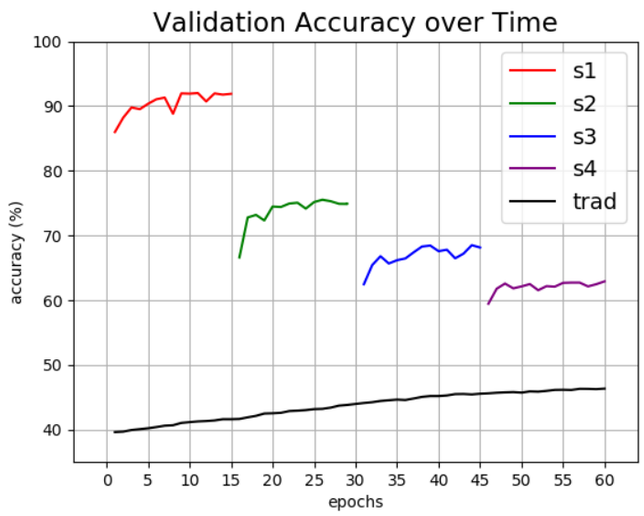

Figure 10: Curriculum learning vs vanilla training

5.3 Curriculum Learning for Age Estimation

Four our age estimation network, ANet, we trained it directly as a multi-class classification problem using the UTKFace dataset [25]. We utilized a novel curriculum learning method by gradually increasing the classification objective difficulty instead of ordering the training data in terms of difficulty [26]. Concretely, ANet will initially be tasked with classifying the images into two age classes: 0-50, 50+. After a period of training, the task difficulty

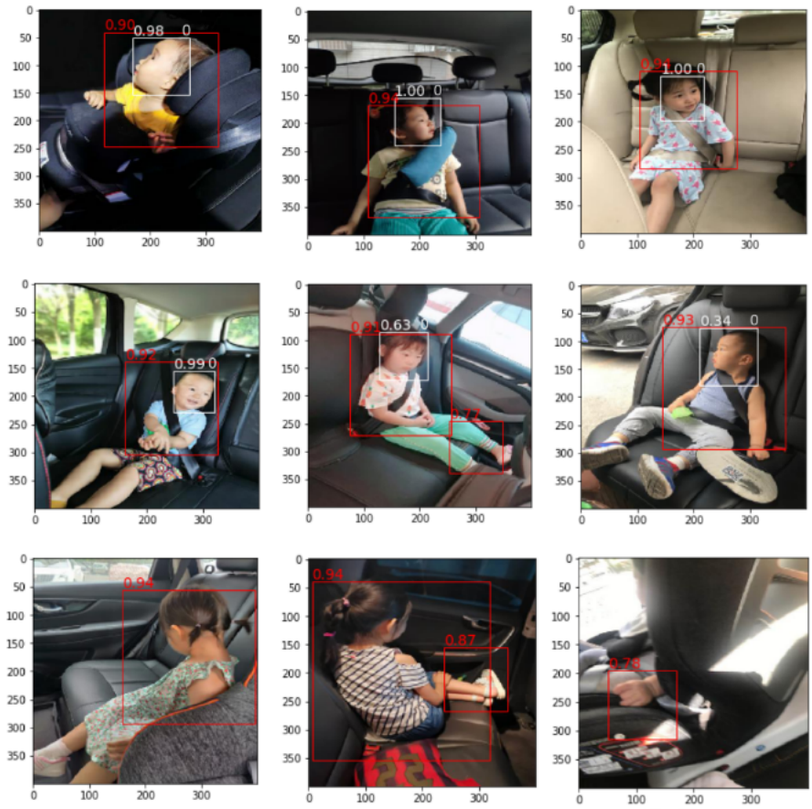

Figure 11: Results demonstration

will be increased by further subdividing objective to four age classes, 0-25, 25-50, 50-75, 75+. The process continues until the desired level of subdivision is reached (in our case, this is 10 age classes, 0-10, 10-20, and so on). Our experiments demonstrate that this method is effective. In Figure 10, *s1-4* refers to the four stages of curriculum learning where we gradually increase the classification objective difficulty, and *trad* refers to the vanilla training method where ANet is trained on a 10-class classification problem from the beginning. In the end, the curriculum-learning based method achieved a validation accuracy of 62%, while the vanilla method only achieved 46%. In addition, for our age-group of interest in the PVH problem, which is 0-10, ANet trained with our curriculum learning method achieved a F1 score of 0.91, while the traditional version only reached 0.66.

6 Conclusion

Some results of our finished pipeline are presented in Figure 11. Our pipeline outputs a red bounding box for any detected children bodies (with the confidence score in the upper-left corner), and a white bounding box for any faces detected (with the confidence score on the upper-left and age-class on the upper-right). The first two rows present the model's outputs when the children's face is clearly visible: our model is able to identify both the body and face, and classify the age. Row three shows our model's performance on images where faces aren't present. Due to our data augmentation techniques, our model can even detect children under extreme occlusion, where only a hand or leg is visible (as shown in the last two images in row 3).

In conclusion, we developed a machine-learning based embedded warning system that helps to mitigate the occurrence of Pediatric Vehicular Heatstroke. Through the Children in Vehicles (CIV) dataset, we trained our multi-task cascaded convolutional neural network and verified its effectiveness on detecting the presence of children. Additionally, we demonstrated the power of our novel curriculum learning method through gradually increasing the

classification subdivision during training on improving the performance of age estimators. In the future, we may work on integrating the 3 second-stage models into a single multi-task network, to further reduce the model size and allow for deployment onto even smaller and cheaper embedded devices.

7 References

[1] J. Null. (2020). No heat stroke, [Online]. Available: https://www.noheatstroke.org/.

[2] D. Diamond. (2014). Cognitive and neurobiological perspectives on why parents lose awareness of children in cars, [Online]. Available: http://psychology.usf.edu/faculty/data/ddiamond/Research_on_Why_Parents_Forget_Children_in_Hot_Cars.pdf.

[3] M. Lynberg. (2018). Heatstroke, NHTSA, [Online]. Available: https://www.nhtsa.gov/campaign/heatstroke.

[4] A. Guard, "Heat related deaths to young children in parked cars: An analysis of 171 fatalities in the united states, 1995-2002," *Injury Prevention*, vol. 11, no. 1, pp. 33–37, Feb. 1, 2005. [Online]. Available: http://ip.bmj.com/cgi/doi/10.1136/ip.2003.004044.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 24, 2017. [Online]. Available: http://dl.acm.org/citation.cfm?doid=3098997.3065386.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556 [cs]*, Apr. 10, 2015. arXiv: 1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385 [cs]*, Dec. 10, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv:1409.4842 [cs]*, Sep. 16, 2014. arXiv: 1409.4842. [Online]. Available: http://arxiv.org/abs/1409.4842.

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *arXiv:1608.06993 [cs]*, Jan. 28, 2018. arXiv: 1608.06993. [Online]. Available: http://arxiv.org/abs/1608.06993.

[10] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," in *arXiv:1704.04861 [cs]*, Apr. 16, 2017. arXiv: 1704.04861. [Online]. Available: http://arxiv.org/abs/1704.04861.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *arXiv:1311.2524 [cs]*, Oct. 22, 2014. arXiv: 1311.2524. [Online]. Available: http://arxiv.org/abs/1311.2524.

[12] R. Girshick, "Fast r-CNN," *arXiv:1504.08083 [cs]*, Sep. 27, 2015. arXiv: 1504.08083. [Online]. Available: http://arxiv.org/abs/1504.08083.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," *arXiv:1506.01497 [cs]*, Jan. 6, 2016. arXiv: 1506.01497. [Online]. Available: http://arxiv.org/abs/1506.01497.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-CNN," *arXiv:1703.06870 [cs]*, Jan. 24, 2018. arXiv: 1703.06870. [Online]. Available: http://arxiv.org/abs/1703.06870.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *arXiv:1506.02640 [cs]*, May 9, 2016. arXiv: 1506.02640. [Online]. Available: http://arxiv.org/abs/1506.02640.

[16] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7553523/.

[17] R. Rothe, R. Timofte, and L. V. Gool, "DEX: Deep EXpectation of apparent age from a single image," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, Chile, Dec. 2015, pp. 252–257. [Online]. Available: http://ieeexplore.ieee.org/document/7406390/.

[18] W. Cao, V. Mirjalili, and S. Raschka, "Rank-consistent ordinal regression for neural networks," *arXiv:1901.07884 [cs, stat]*, Aug. 5, 2019. arXiv: 1901.07884. [Online]. Available: http://arxiv.org/abs/1901.07884.

[19] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, Jun. 2017. arXiv: 1611.01731. [Online]. Available: http://arxiv.org/abs/1611.01731.

[20] Tzutalin, *Tzutalin/labelImg*, original-date: 2015-09-17T01:33:59Z, Aug. 11, 2020. [Online]. Available: https://github.com/tzutalin/labelImg.

[21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," p. 12,

[22] S. van der Walt, S. C. Colbert, and G. Varoquaux, "The NumPy array: A structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, Mar. 2011. [Online]. Available: http://ieeexplore.ieee.org/document/5725236/.

[23] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997. [Online]. Available: http://link.springer.com/10.1023/A:1007379606734.

[24] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv:1609.04747 [cs]*, Jun. 15, 2017. arXiv: 1609.04747. [Online]. Available: http://arxiv.org/abs/1609.04747.

[25] Z. Zhang. (2017). UTKFace, [Online]. Available: https://susanqq.github.io/UTKFace/.

[26] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, Montreal, Quebec, Canada, 2009, pp. 1–8. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1553374.1553380.

8 Acknowledgment

I became aware of the Pediatric Vehicular Heatstroke problem four years ago when I came across an online news article reporting such an incident, where an infant was forgotten in a car for eight hours on a hot summer day. The lifeless body of the infant curled up in the backseat and the image of the devastated father impacted me deeply and lingered in my mind, which motivated me to research about PVH and raise people's awareness of this problem. As a part of my work I wanted to develop a technological solution that would help prevent such tragedies, which culminated in this research project.

I would like to first express my sincere gratitude to my parents. Not only did they provided me with support and positivity during the entire process, they also enlightened me with their philanthropic philosophy, which built my sensitivity towards social issues and allowed me to discover the problem of Pediatric Vehicular Heatstroke.

I would also like to thank the relatives, friends, colleagues, and their children who generously volunteered their time and effort to help with the construction of the Children in Vehicles dataset. Without them this research project would not be possible.

Last but not least, I would like to extend my thanks to my adviser, Mr. Karasik, who provided me with guidance and suggestions throughout the paper-writing process.