Point of care ultrasound for neonatal health

Arijit Patra
arijitp13@gmail.com

Abstract

A major challenge in pre-natal healthcare delivery is the lack of devices and clinicians in several areas of the developing world. While the advent of portable ultrasound machines and more recently, handheld probes, have brought down the capital costs, the shortage of trained manpower is a serious impediment towards ensuring the mitigation of maternal and infant mortality. Diagnosis of pre-natal ultrasound towards several key pre-natal health indicators can be modelled as an image analysis problem amenable to present day state-of-the art deep learning based image and video understanding pipelines. However, deep learning based analysis typically involves memory intensive models and the requirement of significant computational resources, which is a challenging prospect in point-of-care healthcare applications in the developing world. With the advent of portable ultrasound systems, it is increasingly possible to expand the reach of prenatal health diagnosis. To accomplish that, there is a need for lightweight architectures that can perform image analysis tasks without a large memory or computational footprint. We propose a lightweight convolutional architecture for assessment of ultrasound videos, suitable for those acquired using mobile probes or converted from a DICOM standard from portable machines. As exemplar of approach, we validated our pipeline for fetal heart assessment (a first step towards identification of congenital heart defects) inclusive of viewing plane identification and visibility prediction in fetal echocardiography. This was attempted by models using optimised kernel windows and the construction of image representations using salient features from multiple scales with relative feature importance gauged at each of these scales using weighted attention maps for different stages of the convolutional operations. Such a representation is found to improve model performances at significant economization of model size, and has been validated on real-world clinical videos.

1 Introduction

A key aspect of the UN Sustainable Development Goals relates to improving reproductive, maternal, newborn and child health. A primary angle to the improvement of maternal and pre-natal health is the adequate monitoring and assessment of fetal growth and abnormalities, so as to devise prognostic and diagnostic measures in the event of possible adverse outcomes such as both anomalies and congenital diseases, the management and cure for some of which require advanced pre-planning even prior to birth due to the technological and capital requirements involved in the management and redressal of several such birth anomalies. Fetal ultrasound is the primary technique for prenatal health monitoring and diagnosis, with several other modalities being restricted. Particularly, Congenital Heart Diseases (CHDs) are responsible for driving infant mortality with rates being 8 in 1000 live births [4], and is therefore a good case study for assessing the efficacy of automated point-of-care systems for pre-natal healthcare delivery. Despite the universally acknowledged applications for ultrasound, systems of image acquisition continue to be expensive and trained manpower is in short supply. Thus, usage of automated image analysis systems built using machine learning algorithms is a potential avenue for improving fetal health monitoring. In recent years, as deep learning based approaches became popular for image processing applications, the size, computational requirements

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

and complexity of models along with data requirements remained a bottleneck towards deployment for point-of-care applications for medical image analysis, despite rapid development in hardware for acquiring ultrasound scans with the help of portable probes and mobile devices. An important clinical step in fetal heart ultrasound characterisation, and essential for prognosis relevant to detetcion and management of CHDs, is the visibility inference (whether or not the heart is visible in the frame) and the standard viewing plane (4-chamber, 3-Vessel or Left Ventricular Outflow Tract/LVOT) identification. This task has been attempted in the past using CNNs, RNNs [7] and temporal fusion [8] but with models used being extremely heavy, of the order of a few hundred MB and a non-trivial inference cost that makes them difficult to deploy on point-of-care systems in low-power and bandwidth areas. While most deep learning methods rely on end-to-end classifications by the feature learning and aggregation capabilities of convolutional networks, we propose to leverage the presence of specific objects and anatomical features defining a viewing plane at multiple scales through a measure of relevance imposed by progressive attention modules [2]. This self-contained measure of importance of features in input allows the models to train only on the features most relevant to the classes under study at the expense of background, thereby reducing the size of the parameter space for such characterisation. This idea leads us to explore the possibility of using attention layers to improve predictive accuracies of lightweight architectures developed for mobile vision application [2,3]. Such mobile optimised models were then ported to a low cost Xiaomi Redmi handset, and a Samsung handset, both running Android 5.0, and used to analyse loaded echocardiography videos with DICOM to avi conversion.

2 Methodology

We attempt to improve the state of mobile ultrasound interpretation by constructing memory efficient mobile deep learning architectures and augmenting the capacity and classification accuracies of the lightweight models so developed by incorporation of an element of hierarchical prioritization of information in the feature space through the use of stage-wise attention maps in the convolution architecture. The idea is that while the usage of customised convolutional layers that use sets of 1x1 and 3x3 filters, with the former serving to impose separation in the depth level of the feature maps, can reduce model size and computational cost, this comes at a reduction in the number of parameters (not necessarily redundant). Such a reduced parameterization without controlling for parameter importance to network decisions adversely affect performance for the given task. This performance loss is reduced in the presented approach by the use of weighted attention mechanisms, where the input images are partitioned into zones that are subsequently weighed to evaluate their contribution towards the final classwise conditional likelihood for the whole image. Such attention based weighing allows improvement in classification without reliance on extraneous model parameters. The role of attention mechanisms in visual understanding of CNNs have been an area of active research. We attempt to identify spatial cues that are most salient in informing the decisions by the convolutional network on the given input. With the parameter budget being constrained for model efficiency, we draw inspirations from the human mind's ability to extract relevant information from a scene towards forming representative knowledge. This is replicated by having a weighted parameterisation of obtained attention maps so as to magnify the impact of the most relevant features in the input space and subdue the background towards final classification probability distributions obtained at the softmax probability layer (Fig. 1). This is effectively a trainable mobile attention module, and can be used at multiple locations in the architecture. The base architecture is inspired by the aggregation of squeeze and excite modules introduced in [1] by substituting larger kernels with 1x1 kernels in multiple layers and using 1x1 plus 3x3 kernels in alternate layers with the proportion of 3x3 filters gradually increased to account for the complexity of neighborhood fine information in higher levels. Each layer, represented as a set s ϵ {1,...,S}, is developed by a set of 1x1 and 3x3 filters that generate the corresponding feature maps for every member of s as F^s=$\{f_1{}^s, f_2{}^s, \ldots, f_n{}^s\}$. This specific manner of representing feature maps is due to our interpretation that every member of the feature map set, $f_i{}^s$, encodes the activations of spatial location i in layer s (each spatial location i is a square region of a 100 x100 grid overlaid on a 2D feature map, so 1<=i<=n, n=100). With different feature map dimensions across layers, the vector F^s has a variable length dimensions for constituent region based encodings. This is resolved using a linear mapping for each of the three sets of F^s obtained to map them to the dimension of that obtained using the final fully-connected layer F''^s, followed by a dot product evaluation of each member of F^s with F''^s. This rationalisation with respect to the final fully connected layer has an additional effect of capturing the overall global representation of the

p1.png

Figure 1: The overall architecture with attention maps at different stages. This is a representation of the configuration SN-att-2. For SN-att-1, the only attention map is after FC-512. There is a dot product with itself before being converted to the attention weight vector in that case, and this is followed by global concatenation towards creating the attention based representation. As example of the approaches, we demonstrate the results on fetal cardiac view identification, which is the fundamental step for observation of heart abnormalities.

Table 1: Performance of our attention driven models on finding standard fetal heart planes, compared to other lightweigth architectures for the fetal heart data

Classification Accuracy (%)							
Method	4C	3V	LVOT	Non-std./BG	Overall	Size	Inference Speedup
Baseline [1]	85.42	70.14	65.71	80.13	75.35	3.24	1x
ShuffleNet [5]	85.50	69.34	67.21	74.11	74.04	1.95	1.2x
MobileNet [6]	87.05	74.25	66.04	77.60	76.24	1.40	1.14x
SN-att-1 (ours)	86.38	78.20	69.12	84.32	79.51	0.29	1.6x
SN-att-2 (ours)	92.60	79.85	78.34	91.52	85.58	0.34	1.5x

input image as well. To obtain weighted attention over multiple layers, a softmax operation is applied over the region-wise dot products with the final encoding. The attention weights so obtained are assigned to each grid region defined for the attention map, and thus are a measure of the contribution of such a region to the overall loss function, i.e, $a_i{}^s = \exp(<f_i{}^s, F'^s>)/\sum_j^n \exp(<f_i{}^s, F'^s>)$, where <,> represents the dot product operation, 1<=i<=n, n=100 here. The attention weights $\{a_1{}^s, a_2{}^s, \ldots, a_n{}^s\}$ so obtained are used to construct the global weighted attention per feature map $m^s{}_a = \sum_i^n a^s{}_i f^s{}_i$ and concatenated to obtain a final layer M= $[m^1{}_a, m^2{}_a, \ldots, m^S{}_a]$ called the Final Image Attention Map in Fig.1, (S=3 in our case). This is followed by a fully-connected layer FC-4 in Fig.1 with C nodes (C is the number of classes considered, C=4 here) for operations of softmax classification loss functions to obtain a classwise probability map. Thus, in effect, the concatenation of the weighted attention maps is used as a substitute for the fully-connected layer driven global image representation for image classification. This operation ensures that different feature regions at multiple scales of processing in convolutional stages are weighted directly and used to inform the final softmax cross-entropy classifier, instead of just using the fully-connected layer obtained by sequential convolutional operations.

3 Results

We start with a limited number of cardiac screening videos from with varying frame rates and containing one or more of the three views of the fetal heart and some background frames. Videos from 80% of the patients are used for training, and the remaining 20% for test experiments. For training, we split available videos into frames and apply data augmentation by an updown and a top-bottom flipping. Individual frames are cropped into 224 x 224 centred about the heart centre, which was known in the ground truth annotations. The models are trained using a batch size of 25 with a learning rate of 0.001. The base architecture has no pooling layers till the fifth 1x1-3x3 layer module to make feature maps available at a higher resolution to make regional attention proposals optimally informative. We derive our attention maps from layer modules 5, 9 and 14. In the absence of established baselines in prior work for mobile based classification in fetal echocardiography datasets, we compare the results obtained for with a standard SqueezeNet architecture [1] adapted for handling our ultrasound image data, and our base model with attention (SN-att- 1 and SN-att-2 with SN-att-1 aggregating the attention layer from the final fully connected layer and assuming different sections as representative of grid-regions in the prior feature maps) for a classification of visibility and viewing planes in fetal echocardiography images. The attention-based approach yields a notable performance improvement despite a negligible addition to the model size (0.29 MB in baseline without attention vs 0.34 MB in SN-att-2), with the overall baseline SqueezeNet accuracy of 75.35 exceeded by both versions of our attention based architectures (79.51 and 85.58). The original SqueezeNet model adapted for this architecture is a heavier model as well. Additionally, the inclusion of weighted

attention improves performance in case of difficult classes like 3V (78.20 and 79.85 vs 70.14 in baseline) and LVOT (69.12 and 78.34 vs 65.71). This is because the weighted attention model allows enhanced reliance on finegrained discriminative features and relatively ignores less-important features in the classification stages. The strategy to include attention layers from different sections of the network as different sections learn different attributes of the image is proven to enable better aggregation of salient features through the improvement by from SN-att–1 (79.51) to SN-att-2 (85.58). It is worth noting that the average performance accuracy for human sonographers ranges from about 62.5 to 85.7% [9, 10] for this task depending on the number of examinations performed. This falls to 52% (and even to 32.50% for sonographers who have done fewer than 2000 examinations) in cases where a congenital anomaly is present. Thus, our proposed models, SN-att-2 in particular does performs competitively with respect to experienced human sonographer performance despite being lightweight enough for portability. To conclude, the ability of attentive classification to focus on relevant features and diminish the role of the background effectively is reflected in the improved top-1 accuracies listed. Such an improvement without a large model complexity addition is of importance in low-compute environments as in mobiles and EDGE devices in the clinical ultrasound space such as probes and portable machines recently introduced. It is worth considering comparisons with quantization models, direct classification baselines from deeper architectures and attention grids with variable resolutions.

As of now, this work has been attempted with competitive accuracies on actual clinical echocardiography videos acquired from multiple clinics, after conversion from the DICOM standard to standard avi formats, which are them processed in our pipelines (the video preparation and the learning/inference stages are therefore separate here). This conversion, is integrated into our method. As future extension to the validations presented, it would be worthwhile to port the pre-trained models along with integration to support input video streams derived using connected probes, similar to the demonstrations attempted for ultrasound to mobile video conversions using handheld probes by industry players. Thus the processing and diagnosis step can be integrated with the real-time acquisition and the whole pipeline can be used end-to-end, with a possible forward integration to cloud services for later quality checks by qualified physicians located away from patient locations.

Acknowledgements

The experiments were performed on a CPU with an Intel Core i7-4770 processor. The authors gratefully acknowledge Intel for support.

References

[1] Iandola, Forrest N., et al. "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size." arXiv preprint arXiv:1602.07360 (2016).

[2] Seo, Paul Hongsuck, et al. "Hierarchical attention networks." CoRR, abs/1606.02393 (2016).

[3] Jetley, Saumya, et al. "Learn to pay attention." arXiv preprint arXiv:1804.02391 (2018).

[4] N. Archer and N. Manning. Fetal Cardiology. Oxford University Press, 2009

[5] Hluchyj, Michael G., and Mark J. Karol. "Shuffle Net: An application of generalized perfect shuffles to multihop lightwave networks." Journal of Lightwave Technology 9.10 (1991): 1386-1397.

[6] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4510-4520).

[7] Huang, Weilin, et al. "Temporal HeartNet: towards human-level automatic analysis of fetal cardiac screening video." International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2017.

[8] Patra, A., Huang, W., Noble, J. A. (2017). Learning Spatio-Temporal Aggregation for Fetal Heart Analysis in Ultrasound Video. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 276-284). Springer, Cham.

[9] Tegnander E, Eik-Nes SH: The examiner's ultrasound experience has a significant impact on the detection rate of congenital heart defects at the second-trimester fetal examination. Ultrasound Obstet Gynecol 2006;28:8-14.

[10] Hernandez-Andrade, E., Patwardhan, M., Cruz-Lemini, M., Luewan, S. (2017). Early Evaluation of the Fetal Heart. Fetal diagnosis and therapy, 42(3), 161-173.