Learning and Understanding Deep Spatio-Temporal Representations from Free-Hand Fetal Ultrasound Sweeps

Yuan Gao(✉) and J.Alison Noble

Biomedical Image Analysis Group, Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, UK
Yuan.Gao2@eng.ox.ac.uk

Abstract. Identifying structures in nonstandard fetal ultrasound planes is a significant challenge, even for human experts, due to high variability of the anatomies in terms of their appearance, scale and position but important for image interpretation and navigation. In this work, our contribution is three-fold: (i) we model local temporal dynamics of video clips, by applying convolutional LSTMs on the intermediate CNN layers, which learns to detect fetal structures at various scales; (ii) we proposed an attention-gated LSTM, which generates spatio-temporal attention maps showing the intermediate process of structure localisation; and (iii) our approach is end-to-end trainable, and the localisation is achieved in a weakly supervised fashion i.e. with only image-level labels available during training. The proposed attention-mechanism is found to improve the detection performance substantially in terms of classification precision and localisation correctness.

Keywords: Spatial-Temporal Neural Network, · Soft Attention · Weakly Supervised Detection · Non-standard Fetal Scan Planes

1 Introduction

Fetal ultrasound screening requires highly experienced sonographers. This makes it difficult for a wider adoption of clinical ultrasound for pregnancy care, especially in low-and-middle-income settings, where there is a substantial lack of experienced sonographers. Therefore, automated scan plane detection algorithms would greatly assist non-expert examination. However, most published algorithms to date only consider detection of standard scan planes, and treat non-standard scan planes as background. Towards more general recognition and navigation, it is necessary to automate the detection and interpretation of contents in non-standard fetal planes. In this work, we present an automated fine-detection system that not only recognizes the standard scan planes but also non-standard planes. This is a more challenging problem due to the need to accommodate greater variations in imaging quality, and high variability of fetal structures in terms of scale, appearance and location.

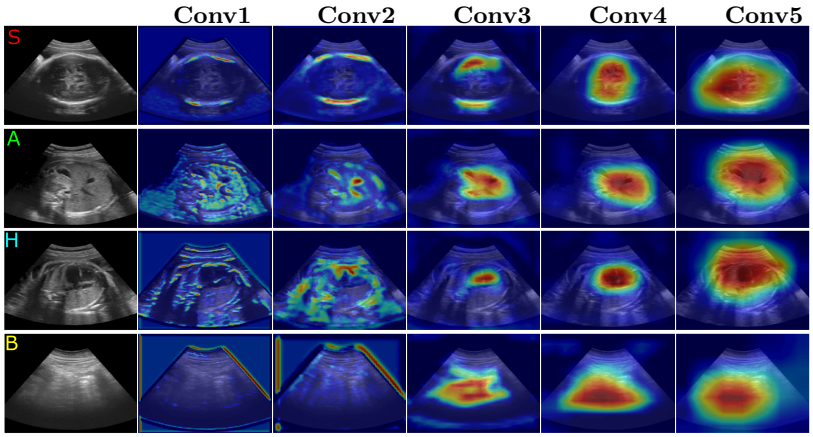


Fig. 1: Class-specific features learned at different layers of a ConvNet. Left-Right: last frame in a sequence, Conv1, Conv2, Conv3, Conv4, Conv5 layers. S: Skull, A: Abdomen, H: Heart, B: Background.

Contributions: In this paper, we introduce a novel attention-gated spatio-temporal neural network (as depicted in Fig.2), trained in a weakly supervised fashion, for detection of key fetal structures of interest (skull, abdomen and heart) in consecutive frames. (1) We apply convolutional LSTMs [1] not solely on the top layer but also on the intermediate features of a convolutional neural network (CNN), to characterise temporal dynamics of features extracted at different scales (as illustrated in Fig.1). (2) We extend the convolutional LSTMs to be attention-gated, regularised by a doubly stochastic mechanism [2] that encourages to fully exploit spatial-temporal information, and more importantly, provides visual evidence of structures localisation. (3) Finally, we investigate the effect of global pooling (average and max) on class activation mapping.

Related Works: Recently, deep learning has become popular for analysing 2D fetal ultrasound sweeps. [3] and [4] have exhibited state-of-the-art performance in still image tasks such as classification or detection. However, such models discard temporal information that for human interpretation we know provides important cues in videos. This is a particularly significant problem for detecting cardiac views. Several recent works [5], [6] and [7] have taken temporal context into account for fetal structure recognition. Chen et al. [5] proposes a recurrent network for detection of standard fetal ultrasound planes, in which they apply LSTMs built on whole image features that completely discard spatial correspondence between frames. Gao et al. [6] proposed a two stream ConvNet, learning spatio-temporal representations to detect the fetal heartbeat in ultrasound videos, which demonstrated a substantial improvement in correctly identifying heart frames. A closely related work to ours is Huang et al. [7] that incorporates a convolutional recurrent layer working at a local region level of the frames. However, and in contrast to our work, the recurrent design is only applied on the coarsest feature map extracted from the CNN. We argue that this design

is more likely to focus on global appearance changes and is not well-suited to capture fine local details, which is nontrivial in our case. In addition, Schlemper et al. [8] proposed a soft attention gated network for improving ultrasound scan plane detection, which is actually an improved SonoNet [5]. However, our work is different in two distinct ways: firstly we propose a general detection framework, not only detecting different anatomies in standard scan planes but also non-standard ones, whilst [5] and [8] only detect standard scan planes and treat non-standard planes as background. Secondly, [8] focuses on aggregating spatial features with spatial attention which did not explore temporal variations and aggregation.

2 Methodology

We present the architecture of our proposed model in Fig.2. Overall, the model consists of a CNN for extracting appearance features from consecutive frames and an Attention Gated LSTM for processing the CNN features recurrently at different locations, exploring temporal variation by selectively attending to different regions of the spatial features maps at different levels.

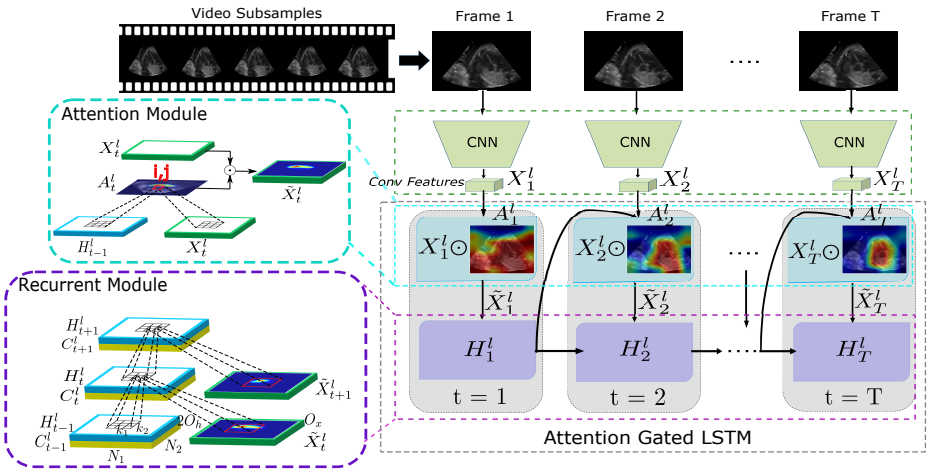


Fig. 2: Network Overview. **CNN feature maps** are fed to the **Attention Module** to generate an attention map at each time step, then the attention-warped (element-wise product) feature maps are fed to the **Recurrent Module** to update the hidden states.

CNN Architecture: Inspired by a VGG very deep architecture [9], we adopt the configuration that increases CNN depth using very small convolution filters stacked with non-linearity injected in between. All convolution layers consist of 3×3 kernels, batch normalization and Rectified Linear Units. The full architecture, using shorthand notation, is 2×C(32,3,1)-MP-2× C(64,3,1)-MP-3×C(128,3,1)-MP-3×C(256,3,1)-MP-3×C(256,3,1)-MP, where $C(d,f,s)$ indicates a convolution layer with d filters of spatial size $f \times f$, applied to the input with stride s. MP represents non-overlapping max-pooling operation with a kernel size of 2×2.

Attention Gated LSTM: The Attention Gated LSTM consists of an attention module and a convolutional LSTM, which are illustrated in Fig.2 showing their inner structure. The max pooled convolutional maps $(X_t^l, ..., X_t^{L-1}, X_t^L)_{(t=1...T)}$, extracted form L layers (L=5 in our CNN) at different time steps in a video, are composed of patterns with strong local correlation over time, and temporal vitiations (e.g. heart beating) tend to be smooth and restricted in a local spatial neighbourhood in successive frames. Therefore, we embed such a prior in our model by replacing the fully connected LSTM gates (i.e. input gates, forget gates etc) with convolution operations. The convolutional kernels (3×3) are chosen to be significantly smaller than the intermediate convolutional map size for computational efficiency.

As noted in Fig.2, instead of directly applying the convolutional maps to the LSTMs, we introduce a soft attention mechanism, conditioned on the convolutional maps X_t^l and their latest hidden state H_{t-1}^l, which learns to focus on salient regions over time:

$$att_t^l = W^l * tanh(W_a^l * X_t^l + U_a^l * H_{t-1}^l + b_a) \tag{1}$$

where again the gating units are characterised by a set of small 2D-convolution kernels: $W_a^l \epsilon \mathbb{R}^{3\times3\times O_x\times O_h}$, $U_a^l \epsilon \mathbb{R}^{3\times3\times O_h\times O_h}$, $W^l \epsilon \mathbb{R}^{3\times3\times O_h\times 1}$ and bias term $b_a \epsilon \mathbb{R}^{O_h}$. *tanh* is non-linearity activation function. The output of this operations is a 2D map from which a normalised attention map is computed through the equation:

$$A_t^l(i,j) = p(att_{ij}|X_t^l, H_{t-1}^l) = \frac{\sigma(att_t^l(i,j))}{\sum_i \sum_j \sigma(att_t^l(i,j))} \tag{2}$$

where $A_t^l(i,j)$ is the element of the attention map in position (i, j), *sigma* is a sigmoid unit. The attention mechanism learns to adaptively guide the LSTMs focusing on the salient(high variant) regions over time and produce the sparse hidden representation $\tilde{X}_t^l$ feeding into the LSTMs, by applying the attention map to the input Conv map with an element-wise product between each channel of the feature maps and the attention map.

Weakly Supervised Localisation: We feed the hidden states of the last time step i.e. H_T^l into the weakly supervised classification layers, which is built with two approaches: (1) 1×1 convolution is applied to mapping the feature maps down to the class score maps, and the score maps are then spatially aggregated using either a Global Max Pooling (GMP) or Global Average Pooling (GAP) operation to obtain categorical scores; (2) We aggregate the feature maps to a feature vector first with global pooling operation, then a dense layer is applied to mapping the feature vector to categorical scores. We then train the proposed model end-to-end by minimizing the objective function:

$$L = -\frac{1}{N}\sum_{n=1}^{N} \alpha_n f_n(S_c(x_n) - log \sum_{k=1}^{K} e^{S_k(X_n)}) + \lambda \sum_{i,j}(1 - \sum_{t=1}^{T} A_t(i,j)) \tag{3}$$

where the first part of L is a focal cross-entropy loss [10]: there are N training sequences x_n and K training classes (K=4), S_k is the k_{th} component in the score

vector $\epsilon\mathbb{R}^K$, and $\mathbf{c}$ is the true class of x_n; α_n is a class-balanced weighting factor set by inverse class frequency; f_n is a focal modulating factor to reduce the loss contribution from well-trained examples and thus focus training on misclassified examples. The second part is a form of doubly stochastic regularization [2] to regularize the learning of spatial-temporal attention. By construction in Eqn.(2), we have $\sum_{i,j} A_t(i,j) = 1$. We also encourage $\sum_t A_t(i,j) \approx 1$ by the doubly stochastic regularization, which can be interpreted as encouraging the model to pay equal attention to every frame and every part of the frame over the course of generation.

3 Experiments and Results

In this section, we evaluate the proposed model jointly for classification and localisation of the fetal structures of interest (skull, abdomen and heart) in video sequences. We compare different weakly supervised classifiers discussed above: AD-GAP (Adaptation with Global Average Pooling), AD-GMP, CAM-GAP and CAM-GMP (Class Activation Mapping with Global Max Pooling). And we also compare against their attention-gated versions: AG-AD-GAP (AG: attention gated), AG-AD-GMP, AG-CAM-GAP and AG-CAM-GMP.

Datasets: Our dataset consisted of 456 fetal ultrasound videos of healthy volunteers, with gestational ages 28 weeks or higher, which have been acquired by a number of experienced obstetricians following a simple protocol i.e. sweeping the ultrasound probe from the maternal cervix to the fundus along the longitudinal axis of the uterus. The videos have been annotated at the frame level by extracting the sequences that contain the anatomy of interest i.e. **Skull**, **Abdomen**, **Heart** and **Background**. The annotation takes both standardised and non-standardised planes into account, and there are significantly more non-standard frames than standard ones because of the simple acquisition protocol. The annotated sequences were sub-sampled with steps conditioned on their length, to create sequences consisting of five consecutive frames. A 5-fold cross-validation is prepared that each fold keeps roughly 20% of the sub-samples for validation and test, the remaining are used for training. The split is made according to the subjects identity to ensure that no samples originating from validation and test subjects were used for training. To evaluate localization, we also annotate bounding-boxes of the fetal structures on the test data. We applied several random on-the-fly data augmentation strategies during training, including (1) cropping square patches at the center of the input frames with a scaling factor randomly chosen between 0.7 to 1, and resize the crops to the size of 224×224 (input resolution); (2) rotation with an angle randomly selected within $\theta = -25°$ to 25°; (3) Random horizontal reflection i.e. flipped the frames in the left-right direction with a probability p=0.5.

Implementation Details: We apply 5 recurrent modules independently on each of the convolution maps extracted from the first to fifth MP (max-pooling) layers. The number of channels of each respective hidden representations are 32, 64, 128, 256 and 256. Five hidden-representations are obtained at each time step. We feed the hidden representations of the last time-step to 5 separate classifiers. Each classifier therefore learns prediction by focusing on only one hidden

representation at a specific scale. The classifier outputs are then averaged to get the final decision. All models are implemented with Tensorflow and trained from scratch (on a Nvidia GeForce GTX 1080Ti) with an Adam optimizer (learning rate: 10^{-4}, β_1=0.5, β_2=0.9 and $\epsilon = 10^{-8}$). λ in Eqn.(3) is set to 1 in our experiments.

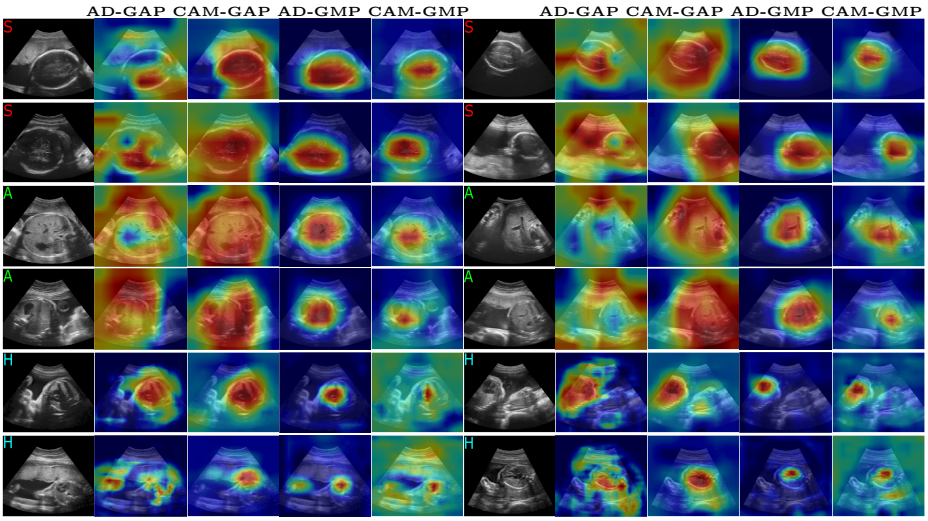


Fig. 3: Examples of the class activation maps obtained from AD-GAP, AD-GMP, CAM-GAP and CAM-GMP. The class activation maps are from Conv5 layer. S: Skull, A: Abdomen, H: Heart.

Class Activation Mapping: Examples of high level (Conv5 layer) class activation maps are illustrated in Fig.3 for different models and different fetal structures. We find that Global Max Pooling (GMP) models, particularly the AD-GMP generates the most discriminative CAMs, localising well the target fetal structures, even in the challenging heart examples with strong acoustic shadow and highly varied appearance and location. While Global Average Pooling (GAP) models perform less superior, the AD-GAP very often failed to capture the anatomy of interest in most cases, and although CAM-GAP localise well the target structures but it tends to over-estimate its extent. We argue that this is because, averaging a feature map is prone to incorporate the global context for the classification, while GMP only accounts for the receptive field of the most discriminative unit i.e. maximally activated neuron.

Attention Analysis: Fig.4 gives examples of spatial-temporal attention captured by the AG-CAM-GMP model at different feature levels. We found that the Conv3 layer captures a rich diversity of temporal dynamics of small anatomies, for instance, the appearing and disappearing of umbilical veins, and the periodic up-down movement of atrioventricular valves. A very interesting heart example (last column), the neural network is struggling to find where the heart is at the

beginning but it is progressively corrected towards end of the video. Likewise, the attention is fairly random initially at Conv4, and because of the extremely large receptive field, Conv5 begin with paying attention on the whole frame, then with incorporation of contextual information, it learns to smoothly and progressively move to the structures of interest, and finally locate them. It should be noted that Conv4 finally focuses more on the anatomies inside the abdomen, such as the heart, stomach bubble and umbilical vein, and Conv5 tends to focus on the whole fetal abdomen.

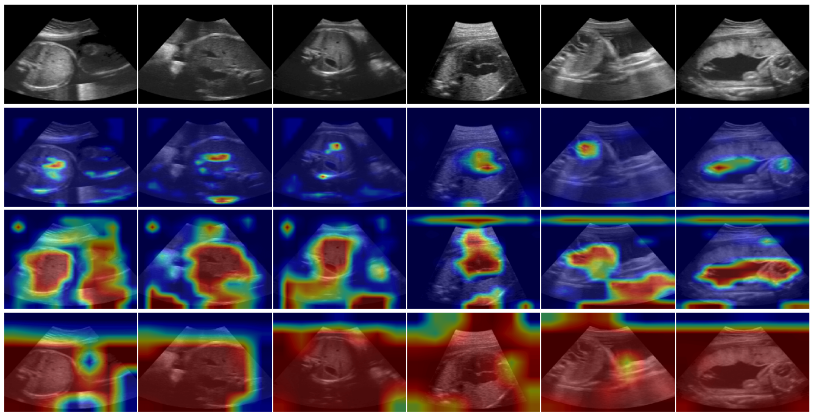
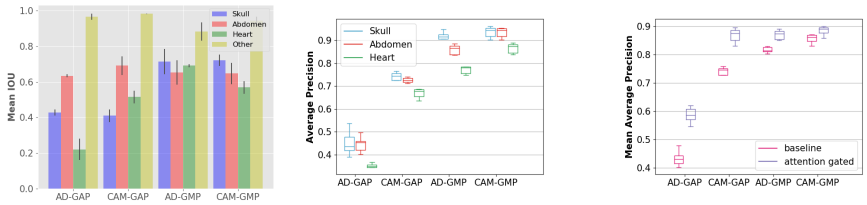


Fig. 4: Examples of multi-scale attention obtained from AG-CAM-GMP. **Row 1**: examples videos. **Row 2 to 4**: attention maps at **Conv3 to Conv5** levels, respectively. Best viewed in Adobe Reader. All videos should play automatically.

Evaluation Localization: We generate a bounding box and its associated anatomy of interest from the class activation maps. We blur Conv4 and Conv5 maps and threshold the lower activations, and then perform a connected component analysis to find the overlapped component, finally we fit a bounding box to the largest connected component. We jointly evaluate the classification and localisation with average precision (AP). We count a correct detection (true positive) if the IOU is above 50% of the maximum achievable IOU of its associated class, and it is a positive prediction with confidence above certain thresholds (ranging from 0.1 to 0.9). Fig.5(b) compares the categorical AP of the baseline models, in which we find the skull and abdomen category achieve comparable performance in each model, and slightly outperform the heart category. The CAM-GMP model achieves the best AP of 0.94, 0.93 and 0.85 (median) in the category of skull, abdomen and heart, respectively. In figure 5(c), we find that attention helps to improve the localisation performance of the baseline models. Particularly, there is a substantial improvement of the mAP (more than 10%) in the GAP models i.e. AD-GAP and CAM-GAP. Although CAMs model already perform very well but there is still a moderate increase of mAP after incorporating attention into the models.



(a) Mean IOU **(b) Average Precision** **(c) Mean Average Precision**

Fig. 5: Quantitative evaluation of localisation performance.

Model Complexity: We have conducted computation complexity studies by considering two metrics: 1) the number of floating-point operations (FLOPs); 2) frame rate (speed) on a single GPU (FPS). FLOPs consist of convolution operations in the backbone (CNN) which is about 3.82 GFLOPs; Recurrent and attention modules is about 1.16GFLOPs and 0.29GFLOPs (here we only consider Conv4 and Conv5 at inference time, as we use conv4 and conv5 CAMs for the localization task); classifiers take a fractional computation that can be neglected compared to the above modules. So in total the proposed baseline model computation is about 4.98 GFLOPs (attention gated models slightly higher at 5.27GFLOPs) that is about one third the computation complexity of VGG16 [9](15.3GFLOPs), and is comparable to SonoNet-32[4]. We also measured the frame rates achieved on a Nvidia Geforce GTX 1080 GPU for classification and localisation combined that base-line models achieve approx. 43.2 FPS, attention-gated models are slightly slower approx. 37.5 FPS. Videos in our study were recorded at 30 FPS so all experimented models achieve real-time performance.

4 Conclusion

We have presented a general framework for detection of multiple fetal structures in free-hand ultrasound videos that it not only learns objects detection from the standardised scan planes but also dominantly from the non-standard cases (especially the abdominal and cardiac planes). Particularly, we proposed a spatial-temporal attention module that can be plug-in any feature level of a ConvNet. As result of multi-scale learning, we have demonstrated the model learns a rich diversity of spatial-temporal patterns at different Conv layers. We also found that the attention helps to improve the localisation performance significantly.

Acknowledgements. We acknowledge the ERC (ERC-ADG-2015 694581, project PULSE) the EPSRC (EP/GO36861/1, EP/MO13774/1) the CSC (DPhil Scholarship No. 201408060107) and the NIHR Biomedical Research Centre funding scheme.

References

1. S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*, 2015.
2. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," In *ICML*, 2015.

3. H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Standard plane localization in fetal ultrasound via domain transferred deep neural networks," *IEEE J. Biomed. Health Inform.*, 2015.
4. C.F. Baumgartner, K. Kamnitsas, S. Smith, L.M. Koch, B. Kainz, and D. Rueckert, "SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound," In *IEEE TMI*, 2017.
5. H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, "Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks," In *MICCAI*, 2015.
6. Y. Gao and J. A. Noble, "Detection and characterization of the fetal heartbeat in free-hand ultrasound sweeps with weakly-supervised twostreams convolutional networks," In *MICCAI*, 2017.
7. W. Huang, C.P. Bridge, J.A. Noble, and A. Zisserman, "Temporal HeartNet: towards human-level automatic analysis of fetal cardiac screening video," In *MICCAI*, 2017.
8. J. Schlemper, O. Oktay, L. Chen, J. Matthew, C. Knight, B. Kainz, B. Glocker, and D. Rueckert, "Attention-Gated Networks for Improving Ultrasound Scan Plane Detection," In *MIDL*, 2018.
9. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," In *ICLR*, 2014.
10. T.Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," In Proc. ICCV, 2017.