Supplementary Material

Knowledge representation and learning of operator clinical workflow from full-length routine fetal ultrasound scan videos

Harshita Sharma[a], Lior Drukker[b], Pierre Chatelain[a], Richard Droste[a], Aris T. Papageorghiou[b], and J. Alison Noble[a]

[a]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom
[b]Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, United Kingdom

In the supplementary material, we describe the experiments and results of the initial exploratory stage of the video description pipeline to determine suitable spatio-temporal network architectures for video clip classification. This was performed to facilitate fully-automatic labelling of full-length ultrasound (US) scan videos.

1 Summary of Spatial and Spatio-temporal Network Architectures

In Sharma et al. (2019), an exploratory spatial analysis was performed for a subset of the US scan video dataset. Three CNN architectures, namely VGG16, VGG19 (Simonyan and Zisserman, 2014) and SonoNet-64 (Baumgartner et al., 2017) (a variant of VGG16), were compared due to their reported good classification performance on natural images and fetal US images, respectively. The analysis was extended to more recent CNN architectures for natural images such as DenseNet-201 (Huang et al., 2017) and MobileNet (Howard et al., 2017). Empirically, SonoNet-64 CNN consistently outperformed the vanilla CNN architectures on spatial subsets and therefore it was selected as the base $2D$ CNN on which to build the spatio-temporal models (Sharma et al., 2019). Also, feature-based and end-to-end spatio-temporal methods were compared in Sharma et al. (2019). However, here we only consider the latter, as these are more conveniently learnt in a single training stage. Furthermore, it has been previously observed that methods based solely on appearance without the use of optical flow could lead to comparable performance for general video classification (Diba et al., 2018). Hence, an additional optical flow input is not considered due to the substantial computational overhead of pre-computing and storing optical flow images for the large-scale US scan video dataset, with each video consisting of thousands of frames.

Table 1 shows all the network architectures considered in the initial exploratory stage. We perform an ablation study, where the individual spatial and spatio-temporal network architectures are compared. Then, we perform model fusions of the individual architectures in different combinations. The notations represent the following layers: CN_2 or CN_3: $2D$ or $3D$ convolution respectively, BN: batch normalisation, MP: max-pooling, $LSTM$: long-short term memory unit, FC: fully connected layer, TD: time-distributed layer, CC: concatenation, SM: softmax, and GAP_2 or GAP_3: global average pooling in $2D$ or $3D$ respectively. The arguments for these elements (if present) are (Feature depth, Kernel size). N_c represents the total number of classes, 12 in the current experiments.

2 Results and Discussion

Quantitative analysis of the investigated models is summarised in Table 2. Highest values of individual metrics from each architecture type are marked in bold. The ablation study confirms that the spatial CNN *SonoNet-64 (FT)* shows superior performance than the other spatial CNN architectures, and the temporally inflated CNN *Sono-2Dt-CNN (RI)* is superior among the two randomly initialised spatio-temporal models. This observation is consistent with the findings

Table 1: Deep Network Architectures

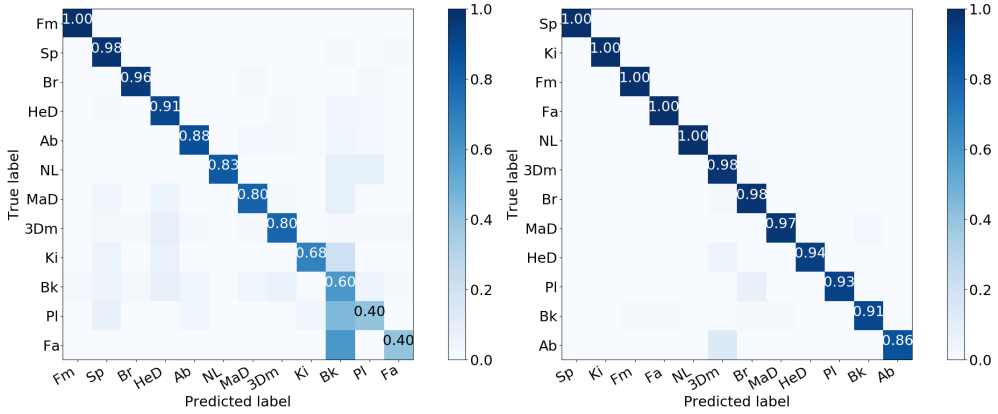
Type	**Network Name**	**Total Params**	**Trainable Params**	**Network Architecture**
Spatial	*SonoNet-64 (PT), SonoNet-64 (RI), SonoNet-64 (FT)* (Baumgartner et al., 2017)	14.87M	14.87M	$2 \times [CN_2(64, 3 \times 3), BN], MP_2(2 \times 2), 2 \times [CN_2(128, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(256, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], CN_2(256, 1 \times 1), BN]$, $CN_2(Nc, 1 \times 1), BN]$, GAP_2, SM
Spatio-temporal	*Sono-2Dt-LRCN (RI)*	11.79M	11.79M	$TD\{2 \times [CN_2(64, 3 \times 3), BN], MP_2(2 \times 2), 2 \times [CN_2(128, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(256, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], MP_2(2 \times 2), CN_2(256, 1 \times 1), BN], CN_2(Nc, 1 \times 1), BN], GAP_2\}, LSTM(256), FC(1024), FC(512), FC(128), FC(N_c), SM$
	Sono-2Dt-CNN (RI) (Sharma et al., 2019)	23.04M	23.04M	$2 \times [CN_3(64, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 2 \times [CN_3(128, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 3 \times [CN_3(256, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 3 \times [CN_3(512, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), [CN_3(256, 1 \times 1 \times 1), BN], [CN_3(N_c, 1 \times 1 \times 1), BN], GAP_3, SM$
Model fusion	*Sono-2D (PT)-2Dt-LRCN (RI), Sono-2D (FT)-2Dt-LRCN (RI)*	27.19M	12.32M	$2 \times [CN_2(64, 3 \times 3), BN], MP_2(2 \times 2), 2 \times [CN_2(128, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(256, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], CN_2(256, 1 \times 1), BN], CN_2(N_c, 1 \times 1), BN], GAP_2, FC(1024) : A$ $TD\{2 \times [CN_2(64, 3 \times 3), BN], MP_2(2 \times 2), 2 \times [CN_2(128, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(256, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], MP_2(2 \times 2), CN_2(256, 1 \times 1), BN], CN_2(N_c, 1 \times 1), BN], GAP_2\}, LSTM(256), FC(1024) : B$ $CC(A, B), FC(512), FC(128), FC(N_c), SM$
	Sono-2D (PT)-2Dt-CNN (RI), Sono-2D (FT)-2Dt-CNN (RI)	39.05M	24.18M	$2 \times [CN_2(64, 3 \times 3), BN], MP_2(2 \times 2), 2 \times [CN_2(128, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(256, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], CN_2(256, 1 \times 1), BN], CN_2(N_c, 1 \times 1), BN], GAP_2, FC(1024) : A$ $2 \times [CN_3(64, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 2 \times [CN_3(128, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 3 \times [CN_3(256, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 3 \times [CN_3(512, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), [CN_3(256, 1 \times 1 \times 1), BN], [CN_3(N_c, 1 \times 1 \times 1), BN], GAP_3, FC(1024) : B$ $CC(A, B), FC(512), FC(128), FC(N_c), SM$
	Sono-2D (PT)-2Dt-LRCN-CNN (RI), Sono-2D (FT)-2Dt-LRCN-CNN (RI)	23.15M	8.28M	$2 \times [CN_2(64, 3 \times 3), BN], MP_2(2 \times 2), 2 \times [CN_2(128, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(256, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(512, 3 \times 3), BN], CN_2(256, 1 \times 1), BN], CN_2(N_c, 1 \times 1), BN], GAP_2, FC(256) : A$ $TD\{2 \times [CN_2(32, 3 \times 3), BN], MP_2(2 \times 2), 2 \times [CN_2(64, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(128, 3 \times 3), BN], MP_2(2 \times 2), 3 \times [CN_2(256, 3 \times 3), BN], MP_2(2 \times 2), CN_2(128, 1 \times 1), BN], CN_2(N_c, 1 \times 1), BN], GAP_2\}, LSTM(128), FC(256) : B$ $2 \times [CN_3(32, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 2 \times [CN_3(64, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 3 \times [CN_3(128, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), 3 \times [CN_3(256, 3 \times 3 \times 3), BN], MP_3(1 \times 2 \times 2), [CN_3(128, 1 \times 1 \times 1), BN], [CN_3(N_c, 1 \times 1 \times 1), BN], GAP_3, FC(256) : C$ $CC(A, B, C), FC(256), FC(128), FC(N_c), SM$

in Sharma et al. (2019). Furthermore, we observe that the performance of individual architectures is lower than the model fusion configurations that leverage the combined power of the transfer-learnt (pre-trained and fine-tuned) spatial networks with the spatio-temporal networks.

The spatial architectures provide a baseline, and as is intuitive, the fine-tuned version outperforms the pre-trained and randomly initialised counterparts. It can be seen that *SonoNet-64 (PT)* shows comparable quantitative performance among the spatial models. However, it should be noted that during testing, sample classes were needed to be readjusted

Table 2: Comparative Analysis of Deep Network Architectures

Type	Network Architecture	P	R	F1	A1	A3
Spatial	*SonoNet-64 (PT)-9 classes*	0.78	0.74	0.72	0.74	0.89
	SonoNet-64 (RI)	0.66	0.68	0.65	0.75	0.88
	SonoNet-64 (FT)	**0.73**	**0.77**	**0.74**	**0.80**	**0.91**
Spatio-temporal	*Sono-2Dt-LRCN (RI)*	0.63	0.68	0.61	0.73	**0.90**
	Sono-2Dt-CNN (RI)	**0.64**	**0.69**	**0.65**	**0.73**	0.87
Model fusion	*Sono-2D (PT)-2Dt-LRCN (RI)*	0.69	0.64	0.64	0.72	0.91
	Sono-2D (FT)-2Dt-LRCN (RI)	0.66	0.74	0.68	0.76	0.92
	Sono-2D (PT)-2Dt-CNN (RI)	0.68	0.64	0.61	0.73	0.87
	Sono-2D (FT)-2Dt-CNN (RI)	0.66	0.75	0.68	0.77	0.93
	Sono-2D (PT)-2Dt-LRCN-CNN (RI)	0.67	0.73	0.68	0.75	0.93
	Sono-2D (FT)-2Dt-LRCN-CNN (RI)	**0.95**	**0.96**	**0.95**	**0.96**	**0.99**



(a) (b)

Figure 1: Confusion matrix predicted *vs.* true label for spatial CNN (left) and spatio-temporal network (right) from the initial exploratory stage.

to map to the original FASP categories (Baumgartner et al., 2017) due to an absence of one-to-one correspondence, leading to 9 resultant classes and ignoring the remaining samples, namely, *Pl*, *MaD* and *3Dm*. Hence, good metrics are obtained for pre-trained *SonoNet-64 (PT)* over only a subset of the considered classes and dataset, and does not fully represent our spatio-temporal semantic segmentation and workflow description problem. Moreover, the remaining categories are found difficult to classify, which is evident from the anatomy-specific results as described below.

Confusion matrix analysis was used to compare spatial and spatio-temporal models. From the confusion matrices in Fig. 1 for the best-performing spatial *SonoNet-64 (FT)* and spatio-temporal *Sono-2D (FT)-2Dt-LRCN-CNN (RI)* networks, we find that most event classes, even when under-represented in the dataset, are more accurately described using $2D + t$ spatio-temporal information than only $2D$ spatial information, suggesting the useful contribution of temporal context for event classification in fetal US scan videos. For the spatial CNN, the result shows higher confusion for *Bk* class, which is understandable as this represents a search process which cannot be described only in the spatial dimension. Classification of *Bk*, along with *Pl*, *Fa*, *Ki*, *NL* and *MaD* increased by introducing the temporal dimension, revealing higher probe motion changes or fetal movements, and indicating difficulty in localisation of structures like kidneys, nose-lips, facial profile and Doppler blood flows of the uterine artery.

References

Baumgartner, C.F., Kamnitsas, K., Matthew, J., Fletcher, T.P., Smith, S., Koch, L.M., Kainz, B., Rueckert, D., 2017. SonoNet: Real-Time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. IEEE Transactions on Medical Imaging 36, 2204–2215. doi:10.1109/TMI.2017.2712367.

Diba, A., Fayyaz, M., Sharma, V., Hossein Karami, A., Mahdi Arzani, M., Yousefzadeh, R., Van Gool, L., 2018. Temporal 3D ConvNets using Temporal Transition Layer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1117–1121.

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 .

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Sharma, H., Droste, R., Chatelain, P., Drukker, L., Papageorghiou, A.T., Noble, J.A., 2019. Spatio-Temporal Partitioning And Description Of Full-Length Routine Fetal Anomaly Ultrasound Scans, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 987–990. doi:10.1109/ISBI.2019.8759149.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR abs/1409.1556.