OBUS – Fetal Presentation Model Architecture

This document describes the architecture used for the development of the Fetal Presentation model. See the GHL OBUS Github Repository [1] for an instantiation of this architecture along with more detail in its Readme file.

Design considerations

Information about the position of fetal anatomical structures, such as the head, abdomen, and limbs relative to the uterus, helps in predicting the fetal orientation. Each video in the blind sweep protocol (see [0.1 OBUS Data Description]), is acquired in a particular direction. For example, the M scan starts at the pubis and sweeps vertically upwards until it reaches the uterine fundus, with the probe indicator facing the maternal right. If the baby were in cephalic position, the head would occur early in the scan and the abdomen, legs, and feet would appear later. Conversely, if the baby were in breech orientation, the feet and legs would be seen first, followed by abdomen and then head. This argues for an architecture that takes the temporal sequence of anatomical structures into account and against an architecture that is order-agnostic, ruling out the Basic Additive Attention module adopted for the Gestational Age model (see [1.2 Gestational Age Model Architectures]).

Two other tendencies may be observed in this regard. The horizontal sweeps (C1, C2, C3, etc.), considered individually, offer less information about whether the baby is in cephalic or non-cephalic orientation. But if the relationship between sweeps were also considered, information could be provided about the cephalic vs non-cephalic question. For example, if the head was seen in C1 and the feet or legs were seen in C3, this would indicate the baby was in breech position. However, the ability to consider inter-sweep relationships would require more complex architectures that involve exam-based training and evaluation. In experimentation, it was found that an architecture that modeled temporal relationships between frames and that was trained at the video level provided sufficient performance as well as simplicity.

Model Architectures

The FP model follows a **Spatial → Temporal aggregation → Classifier** pipeline, with specific choices for the spatial and temporal components.

- **Spatial:** The MobileNet_V2 backbone architecture was found to provide excellent performance and fast throughput.

- **Temporal:** The temporal aggregation component of the architecture chosen was the ConvLSTM model [2], which combines spatial and temporal pattern analysis.
- **Classifier:** The classifier is a fully connected layer with a softmax activation layer.

The architecture block diagram for the Fetal Presentation model is shown in Figure 1.

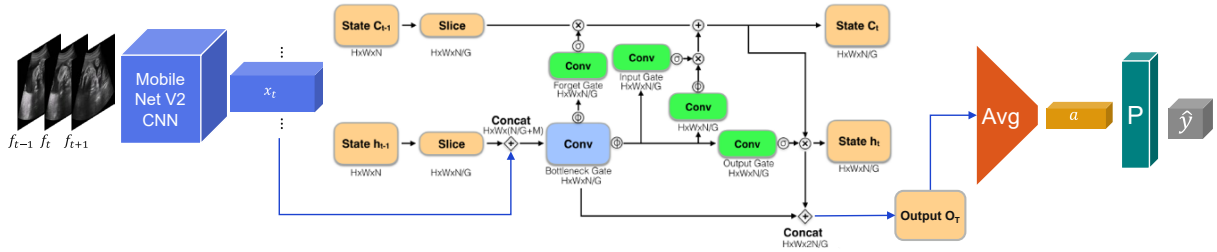

Figure 1. CNN → ConvLSTM → Classifier architecture for FP classification

Training

This model is trained at the video-level on the vertical blind sweeps in the training set. Training the ConvLSTM model requires that a fixed number of frames in every video, at roughly uniform spacing, be selected. At inference time, every vertical blind sweep in an exam folder is evaluated and the output scores are averaged before thresholding. It was observed that performance of the model is optimal when the number of frames selected in inference are the same as the number of frames used when training.

References

[1] "GHL OBUS GitHub Repository," 2025. [Online]. Available: https://github.com/Global-Health-Labs/OBUS-GHL-DEV.

[2] M. Liu, M. Zhu, M. White, Y. Li and D. Kalenichenko, "Looking fast and slow: memory-guided mobile video object detection," *arXiv:1903.10172,* 2019.