OBUS – Gestational Age Model Architecture

This document describes the architecture used for the development of the Gestational Age 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 size of fetal anatomical structures, such as the head, abdomen, and limbs, helps in predicting the age of a fetus. To wit, the biometric method of determining GA uses measurements of these anatomical structures and plugs them into a formula. The difference between the manual biometric method and the automated AI-based method, however, is that the former demands one standard-plane view encompassing the entire structure (*e.g.*, the circumference of the head at a specific cross-sectional plane), whereas the latter may base its prediction on multiple partial views at arbitrary orientations. One might speculate that the AI-based method is less brittle and benefits from an averaging effect over multiple estimates.

This suggests that a frame-based approach might work well for gestational age estimation. While inter-frame spatial correlations might contribute additional performance to the AI-based estimation, experience has shown that a simple order-agnostic weighted average over frames performs well enough, and indeed better than the biometric method on average. The appropriate weighting of the contributions of each frame is a key factor in a frame-based algorithm's performance, for it is clear that many frames do not display any anatomical structures, and indeed, some contain only noise. This also implies that the algorithm need not consider spatial relationships between video sweeps in an exam.

Together, these considerations argue for an architecture that uses an encoder (*e.g.*, CNN) to extract spatial features (*i.e.*, embeddings) from each frame, and a temporal aggregator (*e.g.*, RNN) to aggregate these embeddings into a video feature vector, often called a context vector. A final fully connected layer may then project the context vector to predict the gestational age. To keep neuron activations at a unit scale, the gestational age target is first log-scaled and then z-scaled by the mean-log and std-log values (computed over the training set).

Model Architectures

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

- **Spatial:** Various backbone CNN architectures were trained including EfficientNet_V2_S, EfficientNet_B1, EfficientNet_B0, and MobileNet_V2. The backbones listed here fall in the range between "large and slow but more accurate" to "small and fast but less accurate," the last one being the most deployable on an edge device.
- **Temporal:** A Basic Additive Attention (BAA) [2] model was chosen for the temporal aggregation module. This model is frame-order agnostic, in keeping with the observation above that inter-frame spatial correlations can be ignored.
- **Regressor:** The regressor is a fully connected layer with a linear activation function, trained to output a numeric value as the prediction of the z-scaled-log gestational age estimate.

The model architecture block diagram is shown in Figure 1.

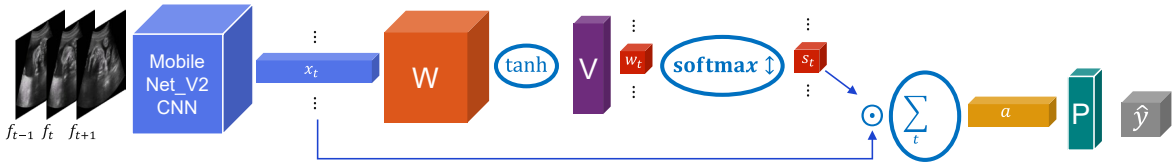

Figure 1. CNN → BAA → Regressor architecture for gestational age estimation.

Training

This model may be trained on the video-level, but memory and batch considerations dictate that a fixed number of frames be sampled (*e.g.*, randomly) during training. At inference time, however, all frames of all videos (given protocol constraints in the number of videos available) in an exam are concatenated and evaluated.

References

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

[2] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," *arXiv:1508.04395v2,* 2016.