Supplement

S.1: Generally: Message passing does not commute with temporal convolutions

Given the feature matrix $X \in \mathcal{R}^{C \times T}$ generated by stacking signals of length T from C different channels, we let A denote the matrix multiplied from the left for message passing (ψ), and $W \in \mathcal{R}^{C \times k}$ be the stack of convolution kernels of length k. The channel-wise time convolution $*_t$ defined for 2d feature matrix is then performed as follows:

$$\mathcal{T}(X) = X *_t W = \begin{bmatrix} X_{1\cdot} * W_{1\cdot} \\ \cdots \\ X_{C\cdot} * W_{C\cdot} \end{bmatrix}$$

where $*$ is the regular 1d convolution along time and entries $[\,]_{i\cdot}$ (e.g. $X_{1\cdot}$) represent the entirety of the ith row vector. We can then calculate $\mathcal{T} \circ \psi = (AX) *_t W$ and $\psi \circ \mathcal{T} = A(X *_t W)$ as follows:

$$\mathcal{T} \circ \psi(X) = (AX) *_t W = \begin{bmatrix} \sum_{j=1}^{C} A_{1j} X_{j\cdot} \\ \cdots \\ \sum_{j=1}^{C} A_{Cj} X_{j\cdot} \end{bmatrix} *_t W \tag{1}$$

$$= \begin{bmatrix} \sum_{j=1}^{C} A_{1j} X_{j\cdot} * W_{1\cdot} \\ \cdots \\ \sum_{j=1}^{C} A_{Cj} X_{j\cdot} * W_{C\cdot} \end{bmatrix}, \tag{2}$$

$$\psi \circ \mathcal{T}(X) = A(X *_t W) = A \begin{bmatrix} X_{1\cdot} * W_{1\cdot} \\ \cdots \\ X_{C\cdot} * W_{C\cdot} \end{bmatrix} \tag{3}$$

$$= \begin{bmatrix} \sum_{j=1}^{C} A_{1j} X_{j\cdot} * W_{j\cdot} \\ \cdots \\ \sum_{j=1}^{C} A_{Cj} X_{j\cdot} * W_{j\cdot} \end{bmatrix}. \tag{4}$$

Thus, in order to have $(AX) *_t W = A(X *_t W)$, one must require that,

$$\sum_{j=1}^{C} A_{ij} X_{j\cdot} * W_{i\cdot} = \sum_{j=1}^{C} A_{ij} X_{j\cdot} * W_{j\cdot}\,, \quad \forall i. \tag{5}$$

or equivalently,

$$\sum_{j=1}^{C} A_{ij} X_{j\cdot} * (W_{i\cdot} - W_{j\cdot}) = 0\,, \quad \forall i. \tag{6}$$

The following example gives a straightforward calculation.[1] Consider the matrices:

$$A = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, X = \begin{bmatrix} 1 & 3 & -1 & -2 \\ -1 & 2 & 1 & 0 \end{bmatrix}, W = \begin{bmatrix} -1 & 2 \\ 3 & 1 \end{bmatrix},$$

so that,

$$(AX) *_t W = \begin{bmatrix} 0.5 & 4 & -0.5 & -2 \\ -0.5 & 3.5 & 0.5 & -1 \end{bmatrix} *_t \begin{bmatrix} -1 & 2 \\ 3 & 1 \end{bmatrix} \tag{7}$$

$$= \begin{bmatrix} 7.5 & -5 & -3.5 \\ 2 & 11 & 0.5 \end{bmatrix} \tag{8}$$

while,

$$A(X *_t W) = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \begin{bmatrix} 5 & -5 & -3 \\ -1 & 7 & 3 \end{bmatrix} \tag{9}$$

$$= \begin{bmatrix} 4.5 & -1.5 & -1.5 \\ 1.5 & 4.5 & 1.5 \end{bmatrix}, \tag{10}$$

thus arriving with $(AX) *_t W \neq A(X *_t W)$.

S.2: The effects of 1st stage training

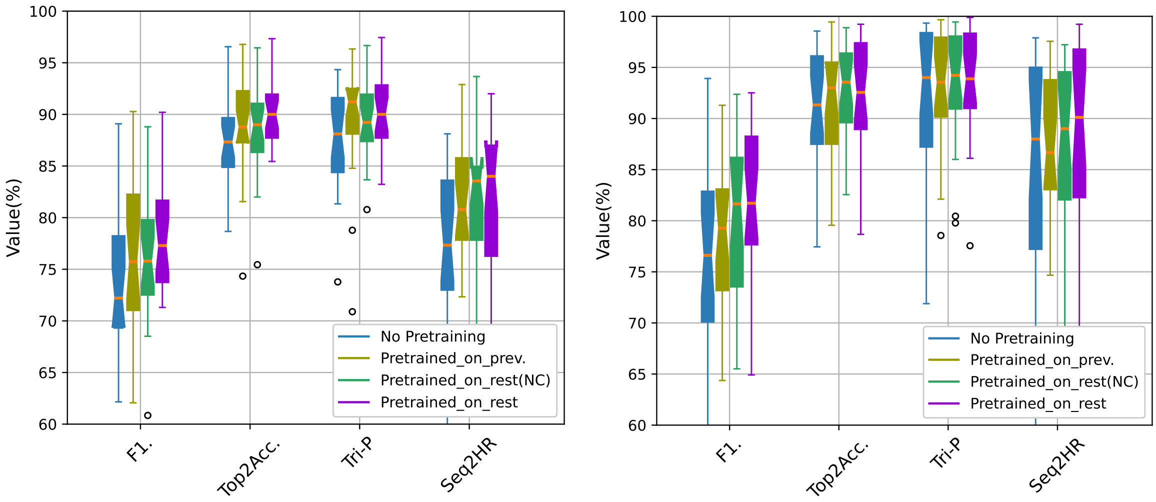

Figure 1: Performance when different pre-training strategies are adopted. Left: Valence. Right: Arousal. NC in the legend stands for no correction being performed for labels before training.

In the configuration of HiSTN-D, we further scrutinize how various choices made during the pretraining stage for the subject-independent experiments could influence the prediction outcomes. The considerations explored here include: 1) omitting data from other subjects entirely and directly training on the restricted 10-second data[2] from the target subject and

[1] The term *convolution* referred to in the neural network setting is actually a *correlation* in standard mathematics terminology, i.e the kernel is not rotated by 180°. In the example, we followed the neural network setting, but one can easily verify that the equation does not hold for either *correlation* or *convolution*.

[2] Training is maintained in two stages under the same setting as in Table *IV*, but the data used in the first stage is replaced with the limited 10-second data.

forecasting on the subsequent 50 seconds; 2) instead of pretraining on data from all remaining subjects, we pretrain only on data from a single subject[3]; 3) adhering to the approach outlined in the subject-independent experiments, but utilizing the subjects' original labels rather than amending them with the most likely one. The outcomes of these experiments are collated and presented in Figure 1.

The figure allows us to make at least three discernible observations as follows:

- Options incorporating a pretraining stage generally surpass those without pretraining across all four metrics evaluated.
- The performances between pretraining with data from the preceding subject and pretraining with data from remaining subjects (without using label correction) are comparable across most metrics. However, the latter choice demonstrated a significantly higher average Seq2HR score for both Valence and Arousal predictions.
- The experimental configuration (pretraining on the remaining data and amending the self-reported score to the most likely score from the empirical distribution), as implemented in our subject-independent experiments, exhibited superior overall performance with the highest mean test value. The only exceptions were the Tri-P value in Valence prediction, and the Top2-Acc. and Tri-P values in Arousal prediction, where it ranked second-best.

[3] We used the subject whose index immediately precedes the selected test subject in the dataset.