Framelet Based Dual Hypergraph Neural Networks for Student Engagement Prediction

Ming Li[1], **Jiandong Shi**[2]

[1]Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, China

[2]Department of Computer Science, Zhejiang Normal University, Jinhua, China

mingli@zjnu.edu.cn, shijiandong@zjnu.edu.cn

Abstract

In this short (**work-in-progres**s) paper, we focus on the critical task of predicting student engagement. We introduce an innovative framelet transform, designed to proficiently convert students' visual features into sets of low-pass and high-pass coefficients. By placing specific emphasis on these coefficients, we create diverse hypergraphs that capture high-order relationships among students at varying scales. Subsequently, we develop dual hypergraph neural networks to effectively learn these hypergraphs, discerning the unique contributions of low-pass and high-pass components. Preliminary experimental findings on a real-world educational dataset highlight the promising potential of our framework in advancing student engagement prediction models.

Proposed Method

Framework Overview

As shown in Figure 1, our framework comprises four key modules. Initially, the constructed undecimated discrete framelet transforms (UDFmT) process students' facial features to obtain frequency-based feature representations. Next, the KNN-based hypergraph generator module constructs hypergraphs that capture complex, higher-order relationships between these features. The hypergraph representation learning module then employs dual hypergraph neural networks to learn from these hypergraphs. Finally, the engagement level classification module uses these representations to classify the engagement level of students into categories such as high or low engagement.

Framelet Transform

Given $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N\times d}$, we next detail the `2D-HaarFrame` that converts efficiently $\boldsymbol{\mathcal{X}}$ to frequency domain with low-pass and high-pass framelet coefficients, $\boldsymbol{\mathcal{X}}_a, \boldsymbol{\mathcal{X}}_{b_i}, i = 1, \ldots, 6$, each of which is also a 2D coefficient matrix in $\mathbb{R}^{N\times d}$, where the `2D-HaarFrame` is determined by a filter bank $\texttt{DHF}_2 = \{a, b_1, \ldots, b_6\}$ of 2D filters. The filter a is the Haar low-pass filter while the other 6 filters are high-pass filters.

A 2D filter (mask) $h = \{h(k)\}_{k\in\mathbb{Z}^2} : \mathbb{Z}^2 \to \mathbb{R}$ is a sequence of filter taps (real/complex numbers) on $\mathbb{Z}^2$. By $\boldsymbol{\delta}$ we denote *the Dirac sequence* such that $\boldsymbol{\delta}(0) = 1$ and $\boldsymbol{\delta}(k) = 0$ for all $k = (k_1, k_2) \in \mathbb{Z}^2\backslash\{0\}$. For $\gamma = (\gamma_1, \gamma_2) \in \mathbb{Z}^2$, we also use the notation $\boldsymbol{\delta}_\gamma$ to stand for the sequence $\boldsymbol{\delta}(\cdot - \gamma)$, i.e., $\boldsymbol{\delta}_\gamma(\gamma) = 1$ and $\boldsymbol{\delta}_\gamma(k) = 0$ for all $k \in \mathbb{Z}^2\backslash\{\gamma\}$. For filters $a, b_1, \ldots, b_L$, we say that a filter bank $\{a; b_1, \ldots, b_L\}$ is a (2-dimensional dyadic) *framelet filter bank* if

$$\sum_{k\in\mathbb{Z}^3} a(\gamma + 2k)\overline{a(n + \gamma + 2k)} + \sum_{i=1}^{L}\sum_{k\in\mathbb{Z}^2} b_i(\gamma + 2k)\overline{b_i(n + \gamma + 2k)} = \tfrac{1}{4}\boldsymbol{\delta}(n), \quad (1)$$

For all $\gamma \in \{0, 1\}^2$ and for all $n \in \mathbb{Z}^2$. Note that

$$\{0, 1\}^2 = \{(0, 0), (1, 0), (0, 1), (1, 1)\} = [0, 1]^2 \cap \mathbb{Z}^2$$

is the set of 8 vertex points in the unit cube $[0, 1]^2$. The filter a is typically a *lowpass filter* satisfying $\sum_k a(k) = 1$ while b_i's are the *highpass filters* satisfying $\sum_k b_i(k) = 0$. Such a filter bank $\{a; b_1, \ldots, b_L\}$ corresponds to a *framelet system* $\{\varphi; \psi_1, \ldots, \psi_L\}$ through refinement relations.

Now we construct a 2D directional Haar filter bank $\texttt{DHF}_2 = \{a, b_1, \ldots, b_6\}$ that satisfies Eq. (1). Consider

$$a^H := \frac{1}{4}\left(\boldsymbol{\delta}_{(0,0)} + \boldsymbol{\delta}_{(0,1)} + \boldsymbol{\delta}_{(1,0)} + \boldsymbol{\delta}_{(1,1)}\right)$$

to be the 2-dimensional Haar low-pass filter. Now, for any two different vertex points γ_1, γ_2 in the unit cube $[0, 1]^2$, we place $+\frac{1}{4}, -\frac{1}{4}$ at each of these two vertices, respectively, and the corresponding high-pass filter is given by $\frac{1}{4}(\boldsymbol{\delta}_{\gamma_1} - \boldsymbol{\delta}_{\gamma_2})$. Collecting all such filters, we have the set $\{b_1, \ldots, b_6\} := \{\frac{1}{8}(\boldsymbol{\delta}_{\gamma_1} - \boldsymbol{\delta}_{\gamma_2}) : \gamma_1, \gamma_2 \in \{0, 1\}^2 \text{ and } \gamma_1 < \gamma_2\}$ of highpass filters. Here $\gamma_1 < \gamma_2$ is understood in the sense of lexicographical order. Then we have in total $L = \binom{2^2}{2} = 6$ high-pass filters. $\texttt{DHF}_2 = \{a^H, b_1, \ldots, b_6\}$ is a tight framelet filter bank such that all the highpass filters $b_1, \ldots, b_6$ have only two taps and exhibit 4 directions in dimension 2. We remark that such types of filter banks exist any dimension $d \geq 1$. In particular, for $d = 1$, the tight framelet filter bank is just the standard Haar orthogonal wavelet filter bank $\text{DHF}_1 := \{a^H, b\}$ with $a^H = \frac{1}{2}(\boldsymbol{\delta}_0 + \boldsymbol{\delta}_1)$ and $b = \frac{1}{2}(\boldsymbol{\delta}_0 - \boldsymbol{\delta}_1)$.

In practice, we employ the **UDFmT** (undecimated discrete framelet transforms) for the decomposition and reconstruction of a 2D matrix. Now we discuss the decomposition and reconstruction of the 2D-matrix $\boldsymbol{\mathcal{X}}$ using our `2D-HaarFrame`. For a 2D filter h, we denote $\boldsymbol{\mathcal{X}}_h$ the (circular) convolution of $\boldsymbol{\mathcal{X}}$ with the 2D filter h, i.e., $\boldsymbol{\mathcal{X}}_h := \boldsymbol{\mathcal{X}} \star h$ with $(k = (k_1, k_2), k' = (k'_1, k'_2) \in \mathbb{Z}^2)$ $\boldsymbol{\mathcal{X}}_h(k) :=$

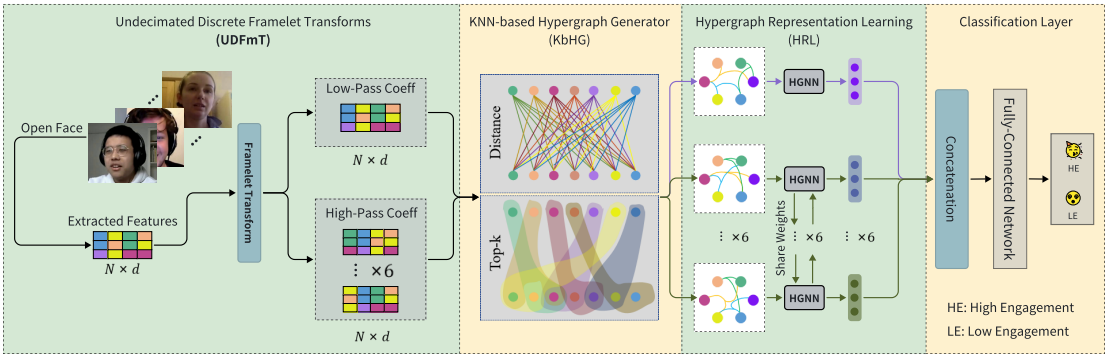

Figure 1: Schematic of the proposed framework.

$\sum_{k' \in \mathbb{Z}^2} \widetilde{\boldsymbol{\mathcal{X}}}(k'-k) \cdot h(k')$, where the above $\widetilde{\boldsymbol{\mathcal{X}}}$ is considered as the periodic extension of $\boldsymbol{\mathcal{X}}$. Note that $\boldsymbol{\mathcal{X}}_h \in \mathbb{R}^{N \times d}$ is a 2D matrix. Consequently, using the filter bank $\mathtt{DHF}_2$, we can decompose $\boldsymbol{\mathcal{X}}$ to 1 low-pass framelet coefficient matrix $\boldsymbol{\mathcal{X}}_a$ and 6 high-pass framelet coefficient matrices $\boldsymbol{\mathcal{X}}_{b_i}, i = 1, \ldots, 6$. The decomposition set $\{\boldsymbol{\mathcal{X}}_a, \boldsymbol{\mathcal{X}}_{b_i}, i = 1, \ldots, 6\}$ of 2D matrices can be used to reconstruct $\boldsymbol{\mathcal{X}}$ perfectly through $\boldsymbol{\mathcal{X}}_a \star \bar{a} + \sum_{i=1}^{6} \boldsymbol{\mathcal{X}}_{b_i} \star \bar{b}_i = \boldsymbol{\mathcal{X}}$, where for a filter h, the filter $\bar{h}$ is defined as $\bar{h}(k) = h(-k), k \in \mathbb{Z}^2$. The set $\{\boldsymbol{\mathcal{X}}_a, \boldsymbol{\mathcal{X}}_{b_i}, i = 1, \ldots, 6\}$ is the one-level decomposition of $\boldsymbol{\mathcal{X}}$. For multi-level decomposition, the input $\boldsymbol{\mathcal{X}}$ is then replaced by $\boldsymbol{\mathcal{X}}_a$ and the filter bank are upsampled, iteratively.

Dual Hypergraph Neural Networks

Leveraging the low-pass coefficients $\boldsymbol{\mathcal{X}}_a$ and the set of high-pass coefficients $\boldsymbol{\mathcal{X}}_{b_i}, i = 1, \ldots, 6$, we construct seven hypergraphs $\mathcal{G}_t$, $t = 1, 2, \ldots, 7$ utilizing the K-nearest neighbor (KNN) methodology. This involves arranging the coefficient-based distances between students in ascending order and selecting the nearest k nodes as neighbors to form a hyperedge.

For each hypergraph $\mathcal{G}_t$, we deploy a two-layer hypergraph neural network, as described by (Feng et al. 2019), to extract node-level representations, meaning individual student embeddings. Notably, to optimize the number of trainable parameters, we implement a weight-sharing scheme during the hypergraph learning phase for $\mathcal{G}_t, t = 2, 3, \ldots, 7$. In the final step, we aggregate the embeddings from each hypergraph $\mathcal{G}_t$ and apply a series of fully-connected networks to yield the engagement level predictions.

Experiments

Dataset

To evaluate the effectiveness of our proposed method, we employ the RoomReader [1] dataset (Reverdy et al. 2022) as a benchmark including over 8 hours of video and audio recordings, capturing the interactions of 118 participants across 30 sessions that take place in the online environment of Zoom.

[1] https://sigmedia.tcd.ie/

Data Preprocessing and Baselines

In our experiments, similar to (Ma et al. 2021), we utilize the normalized eye gaze direction, location of the head, location of 3D landmarks, and facial action units extracted via OpenFace (Baltrusaitis et al. 2018) as the input features. Building upon the work presented in (Reverdy et al. 2022), which provides all the Open-Face features across all sessions in conjunction with multimodal data sources, we conduct experiments on ConvLSTM (Del Duchetto, Baxter, and Hanheide 2020), TEMMA (Chen, Jiang, and Sahli 2020), EnsModel (Thong Huynh et al. 2019), and Bootstrap (Wang et al. 2019) using these features as inputs. We use $k = 3$ in our experiments.

Initial Results

Table 1 presents the results of the performance comparison, which shows clearly that our proposed method outperforms all the baselines in classification accuracy. The mean and standard deviation are obtained based on 5 independent trials.

Table 1: The performance comparison of student engagement prediction accuracy.

Method	**ACC. (%)**
ConvLSTM	76.50 ± 1.85
TEMMA	80.90 ± 2.47
EnsModel	75.30 ± 3.50
Bootstrap	73.80 ± 3.35
Ours	$\mathbf{85.38 \pm 1.41}$

Further Study on Robustness

To evaluate the robustness of our proposed framework, we have conducted a series of experiments wherein additive white Gaussian noise $\boldsymbol{\mathcal{M}} \sim \mathcal{N}(0, \sigma^2)$ is introduced to the feature matrix $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{N \times d}$. The standard deviation of the noise, σ, is determined by $p(\max(\boldsymbol{\mathcal{X}}) - \min(\boldsymbol{\mathcal{X}}))$, where p is selected from $\{0.01, 0.03, 0.05, 0.08, 0.1\}$. Our results demonstrate the models resilience to these varying levels of noise. In follow-up studies, we will delve into the theoretical underpinnings and conduct further experimental investigations to substantiate the benefits of our framework.

References

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. Openface 2.0: Facial behavior analysis toolkit. In *Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition*, 59–66. IEEE.

Chen, H.; Jiang, D.; and Sahli, H. 2020. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23: 4171–4183.

Del Duchetto, F.; Baxter, P.; and Hanheide, M. 2020. Are you still with me? Continuous engagement assessment from a robot's point of view. *Frontiers in Robotics and AI*, 7: 116.

Feng, Y.; You, H.; Zhang, Z.; Ji, R.; and Gao, Y. 2019. Hypergraph neural networks. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 33, 3558–3565.

Ma, J.; Jiang, X.; Xu, S.; and Qin, X. 2021. Hierarchical Temporal Multi-Instance Learning for Video-based Student Learning Engagement Assessment. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, 2782–2789.

Reverdy, J.; Russell, S. O.; Duquenne, L.; Garaialde, D.; Cowan, B. R.; and Harte, N. 2022. RoomReader: A Multimodal Corpus of Online Multiparty Conversational Interactions. In *Proceedings of the 13th Language Resources and Evaluation Conference*, 2517–2527.

Thong Huynh, V.; Kim, S.-H.; Lee, G.-S.; and Yang, H.-J. 2019. Engagement intensity prediction withfacial behavior features. In *Proceedings of the International Conference on Multimodal Interaction*, 567–571.

Wang, K.; Yang, J.; Guo, D.; Zhang, K.; Peng, X.; and Qiao, Y. 2019. Bootstrap model ensemble and rank loss for engagement intensity regression. In *Proceedings of the International Conference on Multimodal Interaction*, 551–556.