Neural Architecture Searching for Facial Attributes-based Depression Recognition

Mingzhe Chen\textsuperscript{1}, Xi Xiao\textsuperscript{1}, Bin Zhang\textsuperscript{2,3}, Xinyu Liu\textsuperscript{1}, Runiu Lu\textsuperscript{1}
\textsuperscript{1}Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
\textsuperscript{2}Department of New Networks, Peng Cheng National Laboratory, Shenzhen, China
\textsuperscript{3}Shenzhen Jinzhou Precision Technology Corp., Shenzhen China
cmz19@mails.tsinghua.edu.cn, \{xiao, lurn\}@sz.tsinghua.edu.cn, bin.zhang@pcl.ac.cn, liuxinyu@chinadrill.com

Abstract—Recent studies show that depression can be partially reflected from human facial attributes. Since facial attributes have various data structure and carry different information, existing approaches fail to specifically consider the optimal way to extract depression-related features from each of them, as well as investigates the best fusion strategy. In this paper, we propose to extend Neural Architecture Search (NAS) technique for designing an optimal model for multiple facial attributes-based depression recognition, which can be efficiently and robustly implemented in a small dataset. Our approach first conducts a warmer up step to the feature extractor of each facial attribute, aiming to largely reduce the search space and providing customized architecture, where each feature extractor can be either a Convolution Neural Networks (CNN) or Graph Neural Networks (GNN). Then, we conduct an end-to-end architecture search for all feature extractors and the fusion network, allowing the complementary depression cues to be optimally combined with less redundancy. The experimental results on AVEC 2016 dataset show that the model explored by our approach achieves breakthrough performance with 27\% and 30\% RMSE and MAE improvements over the existing state-of-the-art. In light of these findings, this paper provides solid evidences and a strong baseline for applying NAS to time-series data-based mental health analysis.

I. INTRODUCTION

Major depressive disorder (MDD) is the most common mental illness with around 5.07\% of adults in the world suffering from it [1]. While the standard clinical depression assessment relies on structured interviews by specially trained psychiatrists [2], it is subjective, time-consuming and hard to access. There is convergent evidence suggests that non-verbal facial behaviors provide rich and reliable sources for reflecting human depression status [3], [4] (e.g., depressed patients usually have reduced facial expressions), and they are easily to be recorded in non-invasive ways using portable devices (mobile phone, laptop, etc.). As a result, a large number of recent studies attempt to automatically recognize depression status from subjects’ faces.

Standard face-based approaches [5], [6], [7], [8] predict depression directly from face images/videos. However, in real-world applications, face images are sometimes not accessible due to various ethical and privacy policies. Since an early study [9] show that mid and low-level facial attributes (e.g., Facial Action Units (AUs) and facial landmarks) are informative for depression status, a certain number of recent studies devote to recognize depression from automatic detected facial attributes such as facial landmarks [10], [11], [12], gaze direction [13], [14], facial action units (AUs) [15], [16], [17], and head poses [10]. Besides some of them compute several statistics [18], [19] (e.g., displacement, velocity, acceleration) from facial attributes time-series as the clip-level representation for depression recognition, recent advances in deep learning (e.g., 1D-CNN [16], [17], LSTM [20], attention-based temporal CNN [21], Causal CNN [22], etc.) also have been applied to infer depression from facial attribute time-series, and achieved enhanced results over most hand-crafted approaches.

However, all of these approaches only manually design models (hand-crafted feature extraction or manually designed CNNs) to extract features from facial attributes, and conduct simple fusion strategy to combine depression cues extracted from all attributes, e.g., they conduct standard decision-level fusion or simply concatenate all facial attributes as a joint representation. Since each facial attribute has a unique data structure, existing approaches fail to design a task-specific architecture for each facial attribute’s feature extraction. Moreover, while each facial attribute contains both unique and common cues (also carried by other facial attribute) for depression recognition, these simple fusion strategies can not optimally retain complementary cues and minimize the redundancy from all facial attributes. In other words, existing approaches that manually design networks are not able to optimally extract and combine depression-related features from multiple facial attributes, which would theoretically limit the recognition performance.

In this paper, we address the aforementioned issues by introducing Neural Architecture Search (NAS) technique to explore an optimal model from small depression dataset, for multiple facial attributes-based depression recognition. Instead of conducting frame/short segment-level depression modelling, our approach starts with employing the spectral encoding method [17] to obtain a clip-level facial behavioral representation for each subject, which has been frequently claimed due be more reliable for depression recognition. Then, we propose a novel multi-stream CNN-GCN framework, where each stream is specifically customized to the unique data structure of a specific facial attribute, aiming to learn extract depression-related features from its clip-level representation, while its fusion module selects a best intermediate latent representation.
from each stream and conducting optimal operations for all representations’ fusion. To achieve this optimal network, we propose an end-to-end NAS strategy to jointly search task-specific architectures for all modules with limited number of depression data. The proposed multi-stream CNN-GCN framework is illustrated in Fig. 1. The main contributions and novelties of our approach is summarized as follows.

- We propose a novel multi-stream framework for multiple facial attributes-based automatic depression recognition, where each stream can be either a CNN or a GNN, and the fusion module optimally combines most informative latent representations that are produced by these streams for depression recognition. To the best of our knowledge, this is the first CNN-GCN framework for face-based depression recognition.

- We propose a Neural Architecture Search (NAS) method to search an optimal multi-stream CNN-GCN framework from a depression dataset that only contains 107 training samples, where a novel motion average loss function is proposed to stabilize the searching process. To the best of our knowledge, this is the first work that extends the Neural Architecture Search (NAS) technique to automatic depression analysis.

- The experimental results show that our approach achieved new state-of-the-art results with 27% RMSE and 30% MAE improvements over the previous state-of-the-art method [21], [19].

II. RELATED WORK

A. Facial attributes-based depression recognition

Due to the ethical concern or storage issue, recent depression recognition challenges [23], [24], [25] encourage researchers to recognize depression from automatically detected facial attributes (e.g., AUs, emotions, facial landmarks, etc.). Jaiswal et al. [18] summarize video-level facial attribute time-series into a histogram, and feed it to MLP to predict the target person’s depression level. Yang et al. [19] introduce a novel hand-crafted video descriptor that manually models the dynamics of 2D facial landmarks of each video segment, which is then fed to a CNN for segment-level depression-related feature extraction. Haque et al. [22] use a Causal Convolutional Neural Network (C-CNN) to deep learn sentence-level depression cues from 3D facial landmarks. Du et al. [21] propose a Atrous Residual Temporal Convolutional Network (DepArt-Net) that generates multi-scale contextual features from several low-level visual behaviors, then temporally fuse them through attention mechanism to capture the long-range depression-related cues. Song et al. [16], [17] propose to use Fourier transforms to encode facial attribute time-series (AUs, gazes, and head poses) of a clip into a length-independent spectral representation, incorporating multi-scale temporal information. However, all these approaches manually design networks for depression feature extraction and fusion without considering the task-specific architecture.

B. Neural Architecture search

Neural Architecture Search (NAS) is an AutoML technique that allows to automatic design a task-specific artificial neural network architecture for the target. Early pioneering work [26] treats neural network architecture search as a combinatorial optimization problem on a discrete search space. A controller optimized by reinforcement learning (RL) [26], [27], [28], [29], [30] or evolutionary algorithms [31], [32], [33] iteratively propose plausible network architectures, where validation results obtained by the explored network is used as a reward signal to update the controller, enforcing it to propose a better architecture. While such strategies achieved promising performance, they are time-consuming. To accelerate searching process, ENAS [29] shares network’s parameters for all candidate architectures. It treats the entire search space
as a super computational graph, and the candidate neural architecture can be viewed as a directed acyclic subgraph. By sharing the model weights among all the different subgraphs (candidate architectures), it is possible to avoid training each subgraph completely from scratch. Recently, Liu et al. [34] propose a DARTS model that replaces the discrete searching process with a continuously differentiable strategy, allowing gradient descent-based architecture optimization and resulting in exponentially faster searching speed. However, such continuously differentiable strategy is less likely to explore an optimal architecture [35], [36], [37], [38] compared with RL-based methods.

III. METHODOLOGY

In this section, we present the details of our depression recognition approach. Our approach first converts all facial attribute time-series of each clip to spectral representations, which provides a set of length-independent clip-level facial attribute representations, summarising various long-term facial behaviors of the target person (Sec. III-A). Then, we describe the proposed multi-stream CNN-GCN depression model in Sec. III-B, which aims to learn task-specific and complementary depression cues from the clip-level representation of each facial attribute while optimally combining them for depression recognition. Finally, we propose an end-to-end strategy to search for the target depression model from a small depression dataset (Sec. III-A).

A. Clip-level facial attribute representations encoding

As discussed above, depression status is more reliable to be reflected from long-term facial behaviors. In this sense, the first step of our approach is to summarize a clip-level representation for the target clip. Since the length of recorded face clip can be various for different subjects, the facial attribute time-series of each subject would be variable. To this end, we apply the spectral encoding algorithm [17] to produce the spectral representations with the same size. If we directly conduct the end-to-end searching for all feature extractors and the fusion module, the search space size for the entire model can be denoted as

\[ S = \sum_{n=1}^{N} \prod_{m=1}^{M} S_{m} \in \mathbb{R}^{C_{m} \times K} \]

where \( S_{m} \) is the search space size for the \( m_{th} \) feature extractor, and \( T_{f} \) is the search space size for the fusion module. Then, their spectral representations can be denoted as \( S_{m}^{\text{amp}} \) and \( S_{m}^{\text{pha}} \). The employed candidate operators for each feature extractor and the fusion module are provided in the supplementary material.

B. Multi-stream CNN-GNN depression analysis model

As illustrated in Fig. 1, the proposed model consists of multiple feature extractors and a fusion module. Each feature extractor aims to extract depression-related cues from the produced spectral representations of a specific facial attribute. In particular, we define a feature extractor as either a CNN or a GNN depending on the data structure of the facial attribute. We first categorize facial attributes into two types: (1) facial attributes whose channels do not have clear spatial correlations (e.g., AUs, gaze and head pose); and (2) facial landmarks that have strict spatial distributions. For each facial attribute time-series of the first type, we concatenate its clip-level spectral representations as a multi-channel heatmap \( S = [S_{n}^{\text{amp}}, S_{n}^{\text{pha}}] \in \mathbb{R}^{C_{n} \times K} \), and use a 1D-CNN to deep learn depression-related feature from it. For facial landmark time-series, we represent them as a clip-level graph. Specifically, we concatenate spectral representations of each facial landmark as a single spectral vector, which is represented as a node feature in the graph. Here, nodes that belongs to a certain facial region (e.g., eyes or mouth) are fully connected, and all nodes are connected to the node that describes the nasal root, allowing messages can be exchanged for different facial regions during the GNN processing.

Meanwhile, we also propose a fusion module to optimally combine depression features produced by all feature extractors. It contains a input block and several fusion blocks. The input block consists of \( N \) parallel input layers, each of which takes the best latent feature from a specific feature extractor. Here, we use the average pooling that aligns all latent representations to the same size, which are then fed to several fusion blocks to optimally combine all depression cues. The employed candidate operators for each feature extractor and the fusion module are provided in the supplementary material.

C. Neural Architecture Search for the multi-stream model

To obtain an optimal multi-stream model to predict depression from multiple facial attributes, this sections proposes to automatically search for the optimal architecture of the model under the supervision of depression labels. Our searching strategy is made up of two stages: the warm-up stage and the end-to-end depression model searching stage.

1) Warm-up stage: Let’s assume \( T_{m} \) is the search space size for the \( m_{th} \) feature extractor, and \( T_{f} \) is the search space size for the fusion module. If we directly conduct the end-to-end searching for all feature extractors and the fusion module, the search space size for the entire model can be denoted as

\[ C = T_{f} \times \prod_{m=1}^{M} T_{m} \]
The $C$ is intimating when the search space size $T_m$ for each module is large. To this end, the warm-up stage first conducts a pre-searching for each feature extractor to reduce the search space size from $T_m$ to $\hat{T}_m \ll T_m$. Consequently, the searching complexity of the warm-up stage is $\sum_{m=1}^{M} \times \hat{T}_m$, while the complexity of end-to-end searching is reduced to $T_f \times \prod_{m=1}^{M} \hat{T}_m$. In short, the warm-up stage reduces the depression network searching complexity to

$$C = \sum_{m=1}^{M} T_m + T_f \times \prod_{m=1}^{M} \hat{T}_m$$  \hspace{1cm} (2)$$

For each feature extractor, we learn a Long-short-term-memory network (LSTM) as the controller to sample architectures. The pseudocode of the warm-up stage is demonstrated in the supplementary material.

2) End-to-end architecture searching stage: Once we obtained the reduced search space for each feature extractor, we then jointly search for the final architecture of all feature extractors and the fusion module. To achieve a high depression recognition performance, this joint searching process enforces all feature extractors to be explored to learn unique and complementary depression cues rather than repeatedly learn depression cues that contained in all facial attributes. Specifically, we individually learn a LSTM as the controller to sample architectures for each feature extractor, all of which are jointly optimized with the controller that samples the architecture for the fusion module. This process is demonstrated in Algorithm 1 and Fig. 2.

3) Optimization details: At each training iteration $t$, the controller (LSTM) samples an architecture $A$ based on its current policy $\pi_\theta$. Then, we instantiate a child network $M$ from $A$ and train it from scratch. The performance of the trained child network is evaluated by the Root Mean Square Error (RMSE) between its predictions and the ground-truth

$$E = \sum_{m=1}^{M} \text{RMSE}$$

trained child network is evaluated by the Root Mean Square Error (RMSE) between its predictions and the ground-truth

$$E = \sum_{m=1}^{M} \text{RMSE}$$

where $M$ and $\{A^m\}$ stands for the architectures of the fusion module and each feature extractor. Finally, the loss function for updating the controller is approximated by an average:

$$L(\theta, \theta_t) = \mathbb{E}_{A \sim \pi_\theta} \left[ J_{\text{ppo}}(A, \theta_t, \theta) \right] \approx \frac{1}{N} \sum_{i=1}^{N} J_{\text{ppo}}(A_i, \theta_t, \theta)$$  \hspace{1cm} (5)$$

where $\{A_i\}_{i=1}^{N}$ are the architectures sampled from $\pi_\theta$.

Since the size of the depression recognition dataset is very limited, the trained models of a certain sampled architecture $A$ can have variable performances due to the stochastic factors of the training (e.g., the shuffled training samples or different initial weights). This means that the validation error in a single trial can not accurately represent the generalization capability of the sampled architecture. In principle, we can mitigate this problem using the Monte Carlo method, averaging the performance by training and evaluating $A$ multiple times each time $A$ is sampled. However, the number of trials used for averaging, denoted as $N$, is difficult to determine, and a large $N$ entails significant computational costs. In this sense, we modify the Monte Carlo method with a motion average (line 17 of the Algorithm 1). This process replaces the validation error, i.e. the reward signal, in a single trial with the average results achieved across the whole training process for $A$. The proposed motion average is formulated as:

$$r(A) = \bar{E}_{\text{val}} = \frac{1}{N} \sum_{i=1}^{N} c_{\text{val}}^{A^i}$$  \hspace{1cm} (6)$$

where $c_{\text{val}}^{A^i}$ is the validation loss obtained by the $i_{th}$ instantiated child network of the architecture $A$. Intuitively, the more frequently an architecture is sampled, the more accurate its performance estimate is, allowing to obtain a more accurate ranking of architectures when the policy converges in the later stage.

IV. EXPERIMENT

A. Dataset

In this paper, we evaluate the proposed approach on a widely-used facial attributes-based depression analysis dataset: DAIC-WOZ dataset[40]. The DAIC-WOZ dataset used in the AVEC 2016 depression recognition challenges contains a total of 189 clinical interviews[11]. The interviews are conducted by an animated virtual interviewer called Ellie, ranging from 7 to 33 minutes in length. Each interview session contains multiple audio-visual and verbal recordings of the participants answering Ellie’s questions, as well as self-assessed PHQ-8 scores(0-24) as ground truth labels.

B. Implementation details

Pre-processing: For each clip, we first remove frames whose face detection are failed or have low confidence of the detected face. We then normalize each of its facial attribute time-series by subtracting the median of the time-series (we retain original values for facial landmark time-series to retain spatial information). Meanwhile, during the spectral encoding, we
Algorithm 1 Joint Neural Architecture Search for our multi-stream depression model

Input: Unified search space $\bigcup_m \{ A^m \}$, Training and Validation datasets $D_{\text{train}}, D_{\text{val}}$, single-modal controllers $\{ \pi_{\theta^m} \}$, fusion controller $\pi_{\theta^f}$, Monte-Carlo steps $N$, Motion average tracker $\mathbb{D}$, Max timesteps $T$

Output: Optimal multi-modal fusion network for depression recognition

1: initialize fusion controller parameters $\pi_{\theta^f}$
2: initialize or load warm-up parameters for single-modal controllers $\{ \pi_{\theta^m} \}$
3: initialize motion average tracker $\mathbb{D} \leftarrow \{ \}$
4: denote parameters of all controllers as $\theta = (\{ \theta^m \}, \theta^f)$
5: for $t = 1$ to $T$ do
6: $L(\theta, \theta_t) \leftarrow 0$
7: for $i = 1$ to $N$ do
8: for each modality $m$ do
9: $\pi_{\theta^m}$ samples a single-modal architecture $A^m \sim A^m$ and output embedding $e_i^m$
10: instantiate single-modal networks $\mathcal{M}_i^m$ based on $A_i^m$
11: end for
12: $\pi_{\theta^f}$ samples the fusion architecture $A_f \sim A^{fuse}$ based on $\{ e_i^m \}$
13: instantiate the fusion network $\mathcal{M}_i^f$ based on $A_i^f$
14: train the joint model $(\{ \mathcal{M}_i^m \}, \mathcal{M}_i^f)$ on $D_{\text{train}}$ from scratch
15: evaluate $(\{ \mathcal{M}_i^m \}, \mathcal{M}_i^f)$ on $D_{\text{val}}$ to get validation error $E_{\text{val}}$
16: update $\mathbb{D}$ with $(A_i, E_{\text{val}})$
17: compute and accumulate $L(\theta, \theta_t)$ with $\mathbb{D}[a_i]$ (See formula (5))
18: end for
19: update $\theta$ with $\nabla L(\theta, \theta_t)$
20: end for
21: return $B$
### TABLE I: Comparison between our approach to others on AVEC 2016 dataset.

| Method            | RMSE | MAE |
|-------------------|------|-----|
| Baseline          | 7.13 | 5.88|
| Williamson et al. [41] | 6.45 | 5.33|
| Song et al. [16]  | 6.29 | 5.15|
| Haque et al. [22] | -    | 5.01|
| Du et al. [21]    | 5.78 | 4.61|
| Yang et al. [19]  | 5.39 | 4.72|
| Ours              | 3.96 | 3.23|

**TABLE II: Results achieved by each explored facial extractor and the fusion module.**

| Method          | AUs  | Gaze | Pose | Landmarks | Fusion |
|-----------------|------|------|------|-----------|--------|
| **RMSE**        |      |      |      |           |        |
| Song et al. [16]| 6.32 | 6.36 | 6.18 | -         | 6.29   |
| Du et al. [21]  | 5.88 | 6.21 | 6.32 | 6.02      | 5.78   |
| Ours            | 5.11 | 5.94 | 5.61 | 4.93      | 3.96   |

| **MAE**         |      |      |      |           |        |
| Song et al. [16]| 5.01 | 5.24 | 5.04 | -         | 5.15   |
| Du et al. [21]  | 4.65 | 4.99 | 5.21 | 5.01      | 4.61   |
| Ours            | **4.38** | **4.83** | **4.62** | **3.82** | **3.23** |

**Graph-based facial landmarks** Table III compares the depression recognition results achieved by applying the explored CNN and GNN to process clip-level facial landmark representations. The results show that the explored GNN clearly outperforms the CNN, suggesting that facial landmark time-series is more suitable to be represented as a graph. Moreover, we show that our fusion module is able to combine the latent representations produced from CNNs and GNN, as the best results is achieved by the CNN-GNN model.

**Motion average loss:** Finally, we also evaluate the proposed motion average loss in Fig 4. As we can see, Fig 4, we demonstrate the learning curve of the controller for each time-step, i.e., the average validation error of the architectures sampled at each step. When adopting our motion average loss, the validation loss curve has decreased variance and less fluctuation. When a single validation error is adopted as the reward signal, the variance is consistently larger and eventually converges to a sub-optimal policy compared to the policy obtained by using our motion average loss.

**V. Conclusion**

In this paper, we propose the first Neural Architecture Search approach to jointly explore optimal CNN-GCN feature extractors and a fusion module for predicting depression from multiple facial attributes. The results show that the model explored by our approach can learn superior depression-related features from all facial attributes, and the fusion module can further enhance the performance by combining depression-related supplementary cues from all facial attributes, with extremely large improvements over the existing state-of-the-art (30% RMSE improvements). In summary, our study provides solid evidences and a strong baseline for applying NAS to automatic depression analysis.
[40] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella et al., “The distress analysis interview corpus of human and computer interviews,” in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), 2014, pp. 3123–3128.

[41] J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H. T. Kung, C. Dagli, and T. F. Quatieri, “Detecting depression using vocal, facial and semantic communication cues,” AVEC 2016 - Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, co-located with ACM Multimedia 2016, pp. 11–18, 2016.