Annales Univ. Sci. Budapest., Sect. Comp. 57 (2024) 167–185

ENHANCING APPARENT PERSONALITY TRAIT ANALYSIS WITH CROSS-MODAL EMBEDDINGS

Ádám Fodor, András Lőrincz and Rachid R. Saboundji
(Budapest, Hungary)

Communicated by Péter Burcsi

(Received January 9, 2024; accepted April 26, 2024)

Abstract. Automatic personality trait assessment is essential for high-quality human-machine interactions. Systems capable of human behavior analysis could be used for self-driving cars, medical research, and surveillance, among many others. We present a multimodal deep neural network with a distance learning network extension for apparent personality trait prediction trained on short video recordings and exploiting modality invariant embeddings. Acoustic, visual, and textual information are utilized to reach high-performance solutions in this task. Due to the highly centralized target distribution of the analyzed dataset, the changes in the third digit are relevant. Our proposed method addresses the challenge of under-represented extreme values, achieves 0.0033 MAE average improvement, and shows a clear advantage over the baseline multimodal DNN without the introduced module.

Key words and phrases: Automatic personality perception, multimodal deep neural networks, deep metric learning.
2010 Mathematics Subject Classification: 97R40.
The research was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program. Á. Fodor and R.R. Saboundji were supported by part through grants EFOP-3.6.3-VEKOP-16-2017-00001 and EFOP-3.6.3-VEKOP-16-2017-00002, respectively. A. Lőrincz was supported by the project "Application Domain Specific Highly Reliable IT Solutions" implemented with the support provided by the National Research, Development and Innovation Fund of Hungary and financed under the Thematic Excellence Programme no. 2020-4.1.1.-TKP2020 (National Challenges Subprogramme) funding scheme.

1. Introduction

Prediction of personality traits is an important task since it is useful for predicting decision-making patterns of people with stable personality traits in diverse situations and detecting changes due to, e.g., stress, drinking, drugs, and so on. One of the most studied model to describe personality is the Big Five personality traits [2]. The theory identifies five factors: Extraversion, Neuroticism, Agreeableness, Conscientiousness, and Openness. Each personality trait represents a range bounded by two extremes, e.g., for extraversion, the two ends are extreme extraversion and extreme introversion.

Audio-visual personality trait prediction has become of high-interest [15] due to high-quality databases released in the ChaLearn challenges, i.e., in First Impressions V1 and V2 [10]. In this study, we used the extended and revised dataset (V2). The dataset contains 10,000 video clips extracted from more than 3,000 different YouTube high-definition videos of people mostly facing and speaking to a camera.

Although multimodal systems offer advantages compared to monomodal systems, they raise several challenges as well. For example, one faces the problems of selecting from the modalities to be included in multimodal systems, deriving the architecture to fuse them, and attenuating errors from noisy, missing, or underrepresented data. One specific characteristic of the First Impression V2 database is its unbalanced data distribution with fewer extreme samples. However, these examples have much more significance and have priority in several use cases, including medication.

Multimodal fusion approaches often hardly consider complex intra- and inter-modal dependencies and lack robustness in case of noisy or missing modalities [26]. Due to these challenges, an increasing number of studies were conducted to transfer knowledge across domains or modalities [11, 20]. Embedding methods have been proven useful for overcoming the aforementioned interdependencies. It has been found that similarity and correlation of semantic information retrieved from real data can be represented using deep metric learning in an embedded feature space [9, 8].

Our contributions are listed below:

1. We propose a multimodal deep neural network for the automatic personality perception task. We extract modality-invariant embeddings from multiple information sources with a distance learning network, emphasizing extreme examples and implicitly improving the multimodal fusion process.

2. We extended the Multi-Similarity loss [21] to handle multiple apparent personality trait class labels simultaneously, besides using various input modalities. The problem with non-extreme examples is that one or more modalities contain inadequate information to aid the deep embedding process. To overcome this issue, we modified the sample selection of the online semi-hard mining procedure to emphasize the extreme samples.

3. Although samples having lower or higher personality trait values are less frequent in the database, high-quality prediction of their values is desired in various situations. We show that cross-modal embedding enhances the prediction of the Big Five personality traits in extreme cases.

The paper is organized as follows. Section 2 reviews the related works. The preprocessing steps, baseline, and the proposed method are detailed in Section 3. The experimental setup, dataset introduction, and implementation details are described in Section 4. Our results, together with the discussion are presented in Section 5. We conclude in Section 6.

2. Related works

Multimodal information has been widely used in various domains ranging from semantic indexing, and multimedia event detection to video situation understanding, among many others. To merge such sources of information, fusion strategies have been derived to harness complementary information from single modalities. Such strategies are classified into three categories, model-level fusion, feature-level fusion, and decision-level fusion [27].

Human behavior monitoring and evaluation rely heavily on multimodal information fusion. Busso et al. [1] paired facial expressions with audio information yielding better prediction for emotion recognition. Wimmer, et al. [23] studied feature-level fusion of low-level audio and video description. Contextual long-range information was later leveraged by the introduction of BLSTMs by Wöllmer et al [24]. In contrast, with the emergence of deep learning, more sophisticated methods were adopted, e.g., by Ngiam et al. [16], who suggested a bi-modal deep auto-encoder to extract shared representations from the input modalities. However, these approaches hardly consider complex intra- and inter-modal interactions and lack robustness in case of noisy or missing modalities [26, 16].

Embedding methods have been proven useful for integrating such interdependencies. Han, J. et al. [7] used triplet loss to distill discriminative representations in the speech modality. Tsai et al. [19] proposed a model that factorizes learned multimodal representations into two sets of independent generative and discriminative factors. Recently, Han et al. [8] introduced a novel learning

framework to leverage information from auxiliary modalities for emotion recognition, using triplet loss to produce modality-invariant emotion embedding in a latent space.

There are recent surveys on personality trait detection that can orient the interested reader [4, 10]. Here, we mention the works related to the challenges called ChaLearn: First Impression Challenge. Kaya et al. [13], the winner of the ChaLearn: First Impressions V1 competition, used visual, audio, and scene features in their system trained end-to-end. Kampman et al. [12] performed an ablation study by combining audio, video, and text information in a trimodal stacked CNN architecture. More recently, Zhang et al. [28] studied the feasibility of merging apparent personality and emotion estimations within a single deep neural network in a multi-task learning framework. An apparent problem with this approach is that the standard deviation of the estimations when trained on the ChaLearn First Impression dataset is much narrower than that of the original data. The phenomenon is called the "regression-to-the-mean problem" where extreme values prediction becomes severely constrained. Li et al. [15] considered this issue and proposed a classification-regression model in which the final regression is guided by the learned classification features and introduced a new objective function called Bell loss to ease the aforementioned problem.

3. Methodology

In this work, we propose a multimodal deep neural network that combines features from visual, acoustic, and textual clues to predict apparent Big-Five personality traits using short video clips from the ChaLearn challenges. The pipeline is depicted in Figure 1.

In the case of audio signals we use standard acoustic features that can be generated by OpenSMILE [5], see later. For the visual feed, most of the frames contain redundant information and we subsample the frames. Since annotated transcripts are noisy, we adopt non-contextual word-level representation for capturing the semantic meaning.

We aim to create a shared coordinate space, transforming the audio, video, and text descriptors into a semantically relevant form using a Distance Learning Network (DLN). DLNs typically use either Siamese networks with contrastive loss or triplet networks with a triplet-based loss function. The triplet-based loss functions are designed to encourage positive examples as close as possible to the so-called anchor sample, and negative examples to be separated from each other over a given threshold. In our experiments, we tested both methods and found that triplet networks outperformed Siamese networks. Embedded vector and auxiliary vector are interchangeably used for the outputs of the

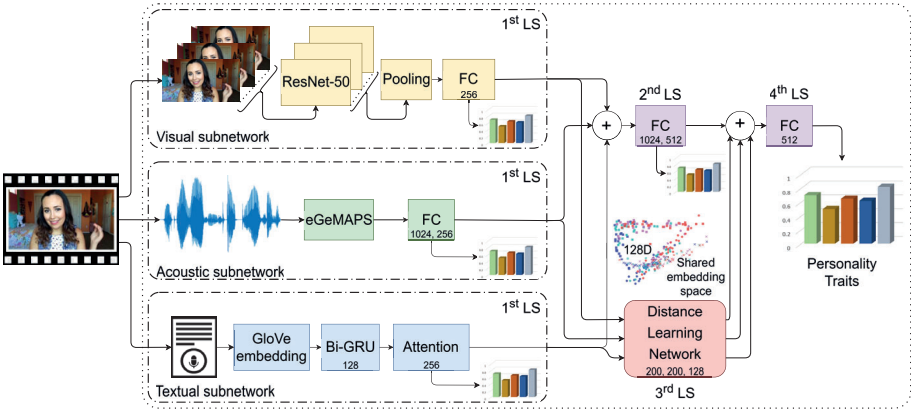

Figure 1. Pipeline of the proposed method for enhanced Big Five personality trait prediction. Visual, acoustic, and textual information are processed with modality-specific subnetworks. The hidden representations are projected into a shared embedding space with a distance learning network (DLN) to exploit mutual information of different information sources implicitly. The shared embedding space of the 128D auxiliary vectors is illustrated by colored circles in 2D. The extracted multimodal hidden representations and the cross-modal embeddings are fused before the final Big Five prediction. The training procedure consists of multiple learning stages (LS). FC: fully-connected, Bi-GRU: bidirectional gated recurrent unit, $\oplus$: concatenation operator. The numbers within blocks indicate the number of hidden units used. Multiple values imply stacked layers.

DLN. Higher precision estimation of the extremes is one of our goals and we expect that multi-modal data enrichment is advantageous in each trait. We use a DNN that combines tri-modal features along with the embedded vectors to predict apparent personality traits from the short video clips.

3.1. Data preprocessing

Audio Features For acoustic features, we used a de-facto standard preset called extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [6]. This feature set contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index, slope V0 features. Furthermore, many statistical functions are applied to these low-level descriptors considering voiced and unvoiced regions, resulting in 88 features for every sample. The audio signals were extracted from the videos using FFmpeg with 44100 sampling frequency. Then, the eGeMAPS were generated through OpenSMILE. Min-max normalization was applied as a preprocessing step to rescale variables into the range [0, 1].

Visual Features We subsampled the video: only 6 frames are selected to reduce the overall complexity and redundancy of successive, similar frames. The choice of 6 is arbitrary and it does not affect the outcome significantly. Pixel values fit into the range of [0, 255]. Images are resized to 140×248 pixels to preserve the original aspect ratio, and then the same random 128×128 pixels spatial crop was applied on all frames of a sample. We employed the same augmentation techniques on every frame of a single clip (with 0.5 probability) during training to preserve the relative similarity between video frames. Data augmentation consists of random flip, random hue (± 0.15), brightness (± 0.2), saturation (between 0.8 and 1.2), and contrast (between 0.8 and 1.2). The augmentations on hue and brightness are additive, while the saturation and contrast are multiplicative. During test and validation time, a center crop was applied. Finally, the frames are scaled between [-1; 1].

Textual Features GloVe uses unsupervised learning to obtain non-contextual vector representations of words. This vector is meant to encode semantic information, such that similar words (e.g., synonyms) have similar embedding vectors. We used pre-trained embeddings (Wikipedia 2014 and Gigaword 5), which capture the overall meaning of a sentence in a relatively lesser amount of memory, and faster than contextual models (like BERT) do. The transcripts are tokenized with SpaCy. All special characters, digits, URLs, and emails are filtered. Every token is converted to its corresponding GloVe vector before feeding it to the textual subnetwork.

3.2. Multimodal information fusion

Visual, acoustic, and textual high-level attributes are combined via a model-fusion approach. Being a regression task, in the first learning stage, the modality-specific subnetworks are trained separately, using ground truth annotations. Hence, they are used as feature extractors, and the parameters of the networks are frozen during further training. In the second learning stage, the tri-modal feature vectors are concatenated and used as the input of a fully-connected network.

Acoustic subnetwork. The 88-dimensional acoustic feature vector $x_A \in \in \mathbb{R}^{\mathrm{N}}$ is the input of the audio subnetwork $f_A : \mathbb{R}^{\mathrm{N}} \to \mathbb{R}^{\mathrm{Q}}$, which is, a fully-connected shallow network with two hidden layers.

Visual subnetwork. Using the video samples $x_V \in \mathbb{R}^{\mathrm{F}\times\mathrm{H}\times\mathrm{W}\times\mathrm{C}}$, where F is the number of frames, H and W are the height and width spatial dimensions, C is the number of channels, a feature extractor $f_V : \mathbb{R}^{\mathrm{F}\times\mathrm{H}\times\mathrm{W}\times\mathrm{C}} \to \mathbb{R}^{\mathrm{Q}}$ is trained. We chose ResNet-50 for our visual backbone. For every frame, a 2048-dimensional feature vector is extracted. Average pooling was applied to the time dimension, followed by a fully-connected layer.

Textual subnetwork. The textual subnetwork input is $x_T \in \mathbb{R}^{\mathrm{K}\times\mathrm{G}}$, where K is the maximum sequence length, G is the dimension of GloVe embeddings. A bidirectional gated recurrent unit (Bi-GRU) with attention mechanism is trained $f_T : \mathbb{R}^{\mathrm{K}\times\mathrm{G}} \to \mathbb{R}^{\mathrm{Q}}$ as a feature extractor.

The x_A audio feature vector, the corresponding x_V RGB frames and x_T GloVe vectors are fed into their modality-specific subnetworks, producing h_A, h_V, $h_T \in \mathbb{R}^{\mathrm{Q}}$ hidden representations, respectively:

$$h_A = f_A(x_A),\ h_V = f_V(x_V),\ h_T = f_T(x_T) \tag{3.1}$$

Let us define $p : \mathbb{R}^{(\cdot)} \to \mathbb{R}^5$, which is a linear mapping function, that estimates the five personality attributes from a given hidden representation. For monomodal subnetworks, the process can be formalized as follows:

$$\hat{y} = p(h_A),\ \hat{y} = p(h_V),\ \hat{y} = p(h_T) \tag{3.2}$$

The network parameters are optimized with Bell loss, following the work of [15]. The shape of the loss function is like an inverted bell and is applied to address the regression-to-the-mean problem [25], which is particularly problematic in our case, where the ground truth scores follow a Gaussian distribution closely. The Bell loss is defined as:

$$\mathcal{L}_{bell} = \frac{1}{5n}\sum_{i=1}^{n}\sum_{j=1}^{5}\gamma\Big(1 - e^{-\frac{(y_{ij}-\hat{y}_{ij})^2}{2\sigma^2}}\Big), \tag{3.3}$$

where n is the number of samples, y_{ij} and $\hat{y}_{ij}$ are the ground truth and prediction of ith sample of jth trait, respectively, σ is the derivation parameter, and γ is a scale parameter. The σ controls the amplitude of variation, and γ makes the loss function consistent with other used loss functions, such as the classical Mean Absolute Error (MAE) and Mean Squared Error (MSE).

$$\mathcal{L}_{mae} = \frac{1}{5n}\sum_{i=1}^{n}\sum_{j=1}^{5}|y_{ij} - \hat{y}_{ij}|,\ \mathcal{L}_{mse} = \frac{1}{5n}\sum_{i=1}^{n}\sum_{j=1}^{5}\big(y_{ij} - \hat{y}_{ij}\big)^2 \tag{3.4}$$

As empirical results showed in [15], the Bell loss has difficulties at the beginning of the optimization and shines at later optimization stages. To avoid the issue, the sum of $\mathcal{L}_{mae}$ and $\mathcal{L}_{mse}$ guide the stochastic gradient descent algorithm in the earlier stages by producing a higher gradient. We trained the modality-specific subnetworks with $\mathcal{L}$, which is the sum of $\mathcal{L}_{mae}$, $\mathcal{L}_{mse}$ and $\mathcal{L}_{bell}$ loss functions introduced in Equation (3.3) and (3.4).

Baseline multimodal network. In the second learning stage, the parameters of the f_A acoustic, f_V visual, and f_T textual subnetworks are not updated.

To leverage the supplementary information of multiple modalities we concatenated the h_A, h_V, and h_T hidden representations and performed model-level fusion. $M_1 : \mathbb{R}^{3Q} \to \mathbb{R}^{O}$ fully-connected shallow network and p is applied to get the personality trait prediction. Formally defined as:

$$\hat{y} = p\Big(M_1\big(h_A \oplus h_V \oplus h_T\big)\Big), \tag{3.5}$$

where $\oplus$ is the concatenation operator.

3.3. Cross-modal deep metric learning

In the following paragraphs, we describe the metric learning framework. We can leverage complementary information from different modalities efficiently using a distance learning network. Using the cross-modal embedding, we make the proposed model more robust to noise, so a more accurate prediction can be achieved. We aim to train a cross-modal DLN $S : \mathbb{R}^{Q} \to \mathbb{R}^{E}$ on the hidden representations of modality-specific nets, which project the multimodal descriptors into a shared coordinate space $\mathbb{R}^{E}$.

$$e_A = S(h_A),\ e_V = S(h_V),\ e_T = S(h_T) \tag{3.6}$$

where S is a DLN, e_A, e_V and e_T are the projected E-dimensional embeddings of h_A, h_V and h_T hidden representations, respectively.

We aim to create a common cross-modal embedding space by transforming tri-modal descriptors in a semantically relevant way. For training the DLN, we choose the current state-of-the-art, triplet-base multi-similarity (MS) loss function [21], which requires an anchor, a positive and a negative example to form positive and negative pairs within a mini-batch. It can jointly measure the self-similarity and relative similarities of a pair, which allows it to collect informative pairs by implementing iterative pair mining and weighting. Deep metric learning requires class labels for training, and MS loss is proposed and tested for only one modality, the single RGB texture.

Triplet generation. Using inputs and the corresponding class labels, we can form triplets $\{e, e^+, e^-\}$. Examples from the same class $\{e, e^+\}$ are determined as positive pairs $\in \mathcal{P}$, as well as samples belonging to different classes $\{e, e^-\}$ are the negative pairs $\in \mathcal{N}$. The Big Five annotations of the First Impressions V2 dataset are continuous variables. We define personality classes in Section 4.4 because it is a database-specific modification.

We applied MS loss, which is defined as a pair weighting problem, and empirical results show that it is superior over other commonly used loss functions, namely the contrastive loss, triplet loss, binomial deviance loss, and lifted structure loss. To compute a cross-modal MS loss, first, the e_A audio, e_V visual, and e_T textual embeddings are combined to form a triple-sized batch of

embeddings denoted as $\{e_A, e_V, e_T\}$, then the similarity metrics are calculated using mixed embeddings of different modalities.

Similarity is defined between two embeddings e_1 and e_2 as the dot product of the vectors considering only the jth personality trait, denoted as $D^j_{e_1,e_2} = \langle S(e_1), S(e_2)\rangle$. MS consists of two parts: mining and weighting. Both schemes are integrated into a single loss function, which is defined as follows:

(3.7)
$$\mathcal{L}_{MS} = \frac{1}{5n}\sum_{i=1}^{n}\sum_{j=1}^{5}\left\{\frac{1}{\alpha}\log\Big[1+\sum_{k\in\mathcal{P}_i^j} e^{-\alpha(D^j_{ik}-\lambda)}\Big] + \frac{1}{\beta}\log\Big[1+\sum_{k\in\mathcal{N}_i^j} e^{\beta(D^j_{ik}-\lambda)}\Big]\right\},$$

where D is the similarity matrix within a triple-size mini-batch, D^j_{ik} is the similarity of two embeddings i and k, $\mathcal{P}^j$ and $\mathcal{N}^j$ are the sets of positive and negative examples considering only the jth trait class labels, respectively. α, β and λ are fixed hyper-parameters.

We calculated a mean, trait-wise multi-similarity loss, considering all 5 target variables per sample within a mini-batch. In the case of non-extreme examples, one or more modalities contain inadequate information to aid the deep embedding process, so we modified the online semi-hard sample mining process to only consider extreme samples as an anchor. In the third learning stage, the S embedding network is trained with trait-wise $\mathcal{L}_{MS}$ with the modified mining procedure. The DLN outputs auxiliary vectors that can help the evaluation due to its specific modality mixing mechanism.

3.4. Fused model

Our method combines the multimodal regression network and the cross-modal distance learning network in the fourth (and final) learning stage. The cross-modal embeddings (e_A, e_V, e_T) are complementary to the hidden representations of the modality-specific subnetworks s and all of them contribute to the final prediction of p.

Model-level fusion is applied, similarly as before: the embeddings are concatenated to the previously fused features, then a $M_2 : \mathbb{R}^{3\mathrm{H}+3\mathrm{E}} \to \mathbb{R}^{\mathrm{O}}$ fully-connected shallow network and p is applied to get the prediction of the Big Five traits. Formally defined as:

(3.8)
$$\hat{y} = p\Big(M_2\Big(M_1\big(h_A \oplus h_V \oplus h_T\big) \oplus e_A \oplus e_V \oplus e_T\Big)\Big),$$

4. Experiments

In the following paragraphs, we introduce the dataset used for the experiments, concretize input and hidden dimensions, and predetermined hyperparameters during the network implementation. Then the evaluation metric, personality trait class definitions, visualization, and the results are presented.

4.1. Database

We used the ChaLearn: First Impressions V2 database for our experiments because it is the largest publicly available in-the-wild dataset in this subfield.

The dataset contains 15 seconds long videos, which are collected automatically. Transcripts of the video clips are generated by a cloud transcription service Rev. The clips are annotated by Amazon Mechanical Turk (AMT) workers using a special interface [17]. Personality annotation followed the Five Factor Model, which consists of Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism, however, the last trait was labeled as Emotional Stability (ES), which is the reverse of Neuroticism and denoted as $\overline{\text{N}}$ in the results section.

They registered annotations using pairwise comparisons, and then they converted the votes to cardinal values by fitting a BTL model with maximum likelihood estimation. Values are scaled, so every video sample has five continuous trait scores between 0 and 1. Each trait represents a range bounded by two extremes. For example, for extraversion, the two polar ends are extreme extraversion and extreme introversion, which can be described with the words “friendly” and “reserved”, respectively. A few examples from the dataset focusing on the extreme poles per trait are depicted in Figure 2.

Creators of this dataset rely on the perception of human subjects watching the videos. It is a different task than evaluating real personality traits with experts, but equally useful in the context of human interaction.

One specialty of this dataset is that the target variables have unbalanced data distribution. The regression-to-the-mean problem is emphasized because the scores follow a Gaussian distribution, and the optimization process likely produces predictions near the mean of ground truth values to minimize the loss. We alleviated this problem with the Bell loss [15], which is similar to the Mean Squared Error, however, it can produce higher gradients when the prediction is closer to the ground truth.

4.2. Experimental setup

Our experiments are conducted with Tensorflow on a single GeForce RTX 2080 Ti GPU. The training process is performed in multiple learning stages.

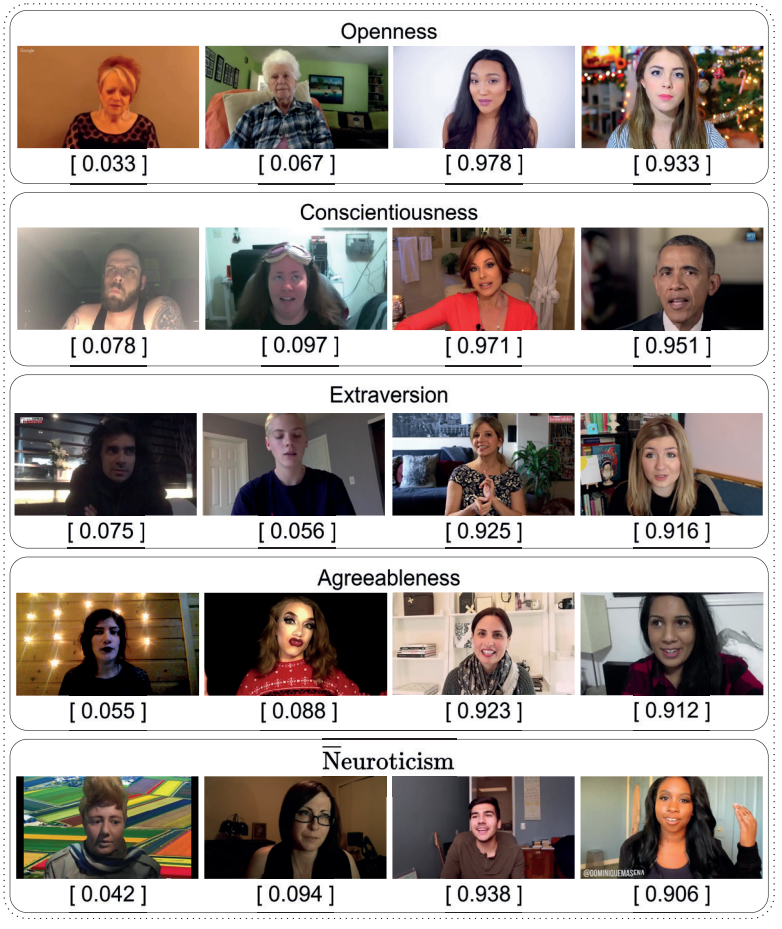

Figure 2. Examples of the First Impression V2 dataset. For each video, the ground truth Big Five scores are provided. For each trait, the first two samples instantiate the low extremes, and the last two examples demonstrate the high extremes of a given trait.

The weights are not modified after a finished stage. We used Adam [14] optimizer with a 0.001 initial learning rate with a polynomial decay schedule throughout all experiments. Following the work of [15], we set the parameters of Bell loss $\sigma = 9$ and $\gamma = 300$. In the first, second, and fourth learning stages, $\mathcal{L}$ was used as the loss function (Section 3.2).

For reduced complexity, we define $Q = 256$ and $O = 512$ in Section 3. All three modality-specific networks produce 256-dimensional feature vectors, and following a concatenation shared dense networks produce 512-dimensional vectors in the baseline and the proposed fused model as well.

For acoustic representation, 88-dimensional eGeMAPS vectors are used ($N = 88$). We fed a mini-batch of 128 vectors to the audio subnetwork and tuned the two fully-connected layers for 100 epochs with early stopping.

After the frame selection (6 frames per clip) and augmentation techniques, $6 \times 128 \times 128 \times 3$ input features are fed to the visual subnetwork ($F = 6$, $H = W = 128$, $C = 3$). We trained it from scratch with a mini-batch of 22 video sequences for 80 epochs. Dropout with a 0.5 rate was applied before the fully-connected layer as an extra regularization.

For semantic word representation, we used 300-dimensional GloVe embeddings. We empirically set the sequence length to 80. After converting every token to its corresponding GloVe vector, an 80×300 matrix is produced for every sample ($K = 80$, $G = 300$). For the textual subnetwork, we used 0.5 for the Bi-GRU input dropout rate. We also applied a simplified attention mechanism [18] and tuned the subnetwork for 50 epochs.

The DLN consists of two fully-connected hidden layers with 200 neurons each and a linear dense output layer with 128 units. Dropout with a 0.5 rate was applied after the first hidden layer. In the third learning stage, we used $\mathcal{L}_{MS}$ as the loss function (Equation (3.7)). We used ReLU as an activation function and Kaiming/He normal initialization, in addition to 0.0005 weight decay in every dense layer, except within the DLN, where weight decay is not considered.

4.3. Evaluation metrics

During the ChaLearn challenge, the "1-Mean Absolute Error" was the performance metric, so many publications employed it. It is defined as follows:

$$R_{acc} = 1 - \frac{1}{5n} \sum_{i=1}^{n} \sum_{j=1}^{5} |y_{ij} - \hat{y}_{ij}|, \tag{4.1}$$

where n is the number of samples, y_{ij} and $\hat{y}_{ij}$ are the ground truth and prediction of ith sample and jth trait.

4.4. Personality trait class definition

Annotation regarding the First Impressions V2 dataset consists of 5 continuous variables. In this work, we aim to differentiate extreme examples from ordinary samples based on the ground truth values. We determine 4 classes per trait, and we are focusing on the two extremes, which can be monitored in various clinical sessions later on: the low-extreme and high-extreme classes, which are labeled as C1 and C4, respectively.

However, in our case, the ground truth follows a Gaussian distribution, and splitting the [0, 1] interval into equal parts would lead us to an undesirably

unbalanced number of extreme samples. To address this issue, we can create more balanced classes by determining the following segmentation thresholds: scores in range $[0, \bar{t} - \sigma_t)$ belong to the low-extreme class (C1), values in range $[\bar{t} - \sigma_t, \bar{t})$ as well as $[\bar{t}, \bar{t} + \sigma_t)$ are labeled as ordinary (C2, C3), and samples between $[\bar{t} + \sigma_t, 1]$ are the high-extremes (C4), where $\bar{t}$ and σ_t is the mean and standard deviation calculated over all training samples of t personality trait. Figure 3 demonstrates the class definitions on the histograms of the train and test sets.

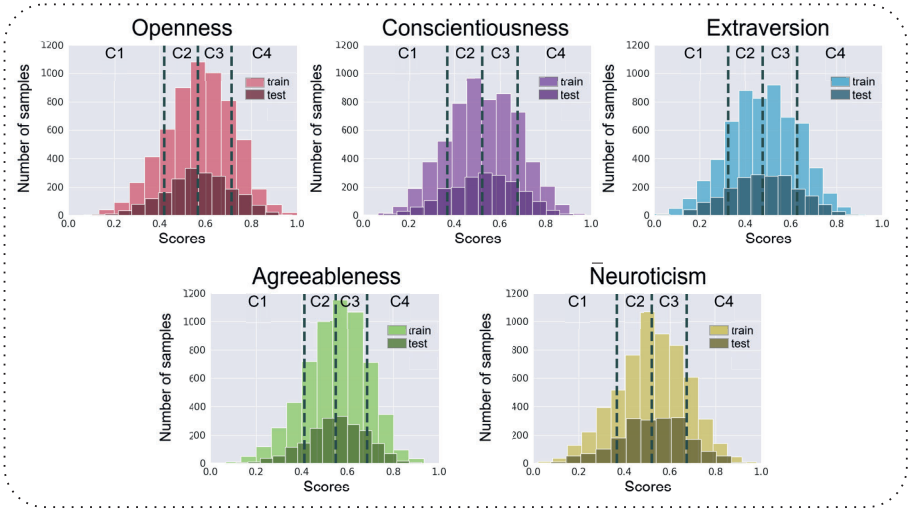

Figure 3. Personality trait class definitions. Continuous ground truth values are segmented into 4 classes. The thresholds are determined using the mean and standard deviation calculated on the train set trait-wise. Samples from C1 and C4 are the low extremes and high extremes, respectively.

4.5. Visualization of cross-modal embeddings

We transformed the acoustic, visual, and textual features into a shared coordinate space with a DLN. Figure 4 shows a two-component Principal Component Analysis (PCA) calculated on the multimodal inputs as visualization, using only the $\overline{\text{N}}$euroticism ground truth values and trait classes within plots.

The test set contains 2000 samples, so considering all three modalities, 6000 embeddings are available. We randomly subsampled to avoid highly overlapped markers and overcrowded visualization, also paying attention to preserving the modality and class balance within the subset: 25 embeddings are selected for every class per modality, so on (a) subplot 300 transformed embeddings are present. The figure shows that even using only two components, the two polar

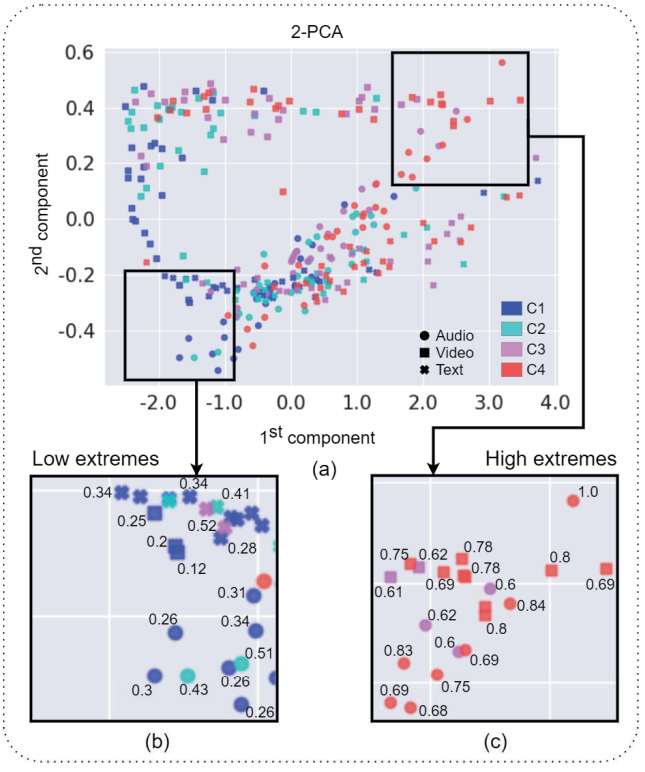

Figure 4. Visualization of 2-component PCA of cross- and multimodal embeddings of the "test" set (a), showing $\overline{\text{N}}$euroticism ground truth values and class labels. The audio, video, and text modalities are drawn with circle, square, and cross, respectively. The four personality classes are represented with colors, where blue is the low extreme (C1), and red is the high extreme class (C4). In (b) and (c), we emphasize embeddings within the two extreme poles of $\overline{\text{N}}$euroticism.

ends of a personality trait are successfully separated. However, there is a continuous transition between trait classes, especially in the case of C2 and C3: the ground truth values are around the mean, and there are hardly perceived or any clues to make these samples more separated using the available inputs.

5. Results

We performed an ablation study with the used modalities to measure the added values of information sources. For the sake of comparison, a prior model obtained directly from the training labels (by averaging) on this dataset was

capable of obtaining close to 0.88 of R_{acc} at test stage [3] due to the highly centralized distribution. In turn, changes in the third digit are relevant. Table 1 indicates that the video modality contains the most information, with an average score of 0.9074. Apparent personality traits can be determined accurately using only a single frame: 0.9056 score over the test set strengthens the statement that trait assignment among human observers can be as fast as 100ms [22]. The bi-modal systems produce a clear performance jump in every single case compared to the monomodal configurations. Furthermore, the "Audio + Video + Text" model performed the expected best result: the different modalities supplement each other.

Thus, we can fairly compare the proposed method to the "Audio + Video + Text" baseline. Table 1 shows that our method performs more superior overall, emphasizing the improvement produced by cross-modal embeddings from 0.9094 to 0.9127.

Input features	O	C	E	A	$\overline{\text{N}}$	Avg
Audio	0.9007	0.8916	0.8947	0.9016	0.8955	0.8968
Scene	0.9048	0.9110	0.9065	0.9065	0.8990	0.9056
Video	0.9065	0.9132	0.9086	0.9072	0.9016	0.9074
Text	0.8900	0.8841	0.8837	0.8982	0.8853	0.8882
Audio + Video	0.9074	0.9143	0.9097	0.9088	0.9041	0.9089
Audio + Text	0.9016	0.8952	0.8958	0.9023	0.8965	0.8983
Text + Video	0.9073	0.9140	0.9105	0.9080	0.9041	0.9088
Audio + Video + Text	0.9069	0.9103	0.9108	0.9108	0.9083	0.9094
Ours	**0.9102**	**0.9154**	**0.9142**	**0.9127**	**0.9112**	**0.9127**

Table 1. Comparison on the R_{acc} performance of the network trained with different data modalities.

We also evaluated the trained system using only extreme samples. We subsampled the test set, so the subset only contained examples from C1 and C4, trait-wise. In Table 2, the "All" column values are produced on the whole test set by the baseline and our method, respectively. In the case of the "Low" and "High" columns, the corresponding personality trait classes are C1 and C4. The results indicate that we can enhance the prediction of high extreme values at the expense of low extreme prediction in most cases. However, by focusing on Conscientiousnes, enhanced quality of both low- and high-extreme predictions can be observed.

6. Conclusions

In this article, we proposed a multimodal deep neural network for the perceived Big Five personality trait prediction, which deals with multimodal data.

	Baseline			Ours		
	All	Low	High	All	Low	High
O	0.9069	**0.8691**	0.8702	**0.9102**	0.8684	**0.8794**
C	0.9103	0.8731	0.8832	**0.9154**	**0.8753**	**0.8891**
E	0.9108	**0.8870**	0.8739	**0.9142**	0.8841	**0.8768**
A	0.9108	**0.8590**	0.8626	**0.9127**	0.8573	**0.8644**
$\overline{\text{N}}$	0.9083	**0.8730**	0.8739	**0.9112**	0.8722	**0.8742**

Table 2. Network performance R_{acc} on all samples, low and high extreme examples. Baseline: Audio + Video + Text. Ours: Audio + Video + Text fused with cross-modal embeddings.

Currently, the largest publicly available dataset is used in these experiments, the ChaLearn: First Impressions V2, and we created embeddings, which are modality-invariant to an extent, to make the different input modalities supplement each other.

An ablation study has demonstrated the added values of different modalities, as well as the proposed extension. We applied a modified multi-similarity constraint over acoustic, visual, and textual representations to implicitly exploit the mutual information. Experiments show that we achieved higher overall prediction accuracy, surpassing the performance of baseline multimodal configurations. Besides, we evaluated the proposed method of extreme examples, which produced the desired results in some cases.

To the best of our knowledge, this is the first work that introduces cross-modal embedding for personality trait prediction. The proposed learning framework is far from perfect. It could be further developed, which is planned for future works. The feature extraction part could be improved to produce more diverse and descriptive representations. Probabilities could be utilized within the triplet constraint to consider the uncertainty around trait class segmentation thresholds properly. The multiple learning phases could be combined to form an end-to-end training process for better useability.

This work is a detailed version of the 13th Joint Conference on Mathematics and Computer Science (MaCS 2020) presentation.

References

[1] **Carlos, B., D. Zhigang, Y. Serdar, B. Murtaza, L. Chulmin, K. Abe, L. Sungbok, N. Ulrich and N. Shrikanth,** Analysis of emotion recognition using facial expressions, speech and multimodal information, *Proceedings of the 6th International Conference on Multimodal Interfaces*, (2014), 205–211.

[2] **Digman, J.M.,** Personality structure: Emergence of the five-factor model, *Annual Review of Psychology*, **41-1** (1990), 417–440.

[3] **Escalante, H.J., H. Kaya, A. Salah, S. Escalera, Y. Güçlütürk, U. Güçlü, X. Baró, I. Guyon, J. C. S. Jacques Junior, M. Madadi, S. Ayache, E. Viegas, F. Gurpinar, A. S. Wicaksana, C. Liem, M. A. J. Van Gerven and R. Van Lier,** Modeling, recognizing, and explaining apparent personality from videos, *IEEE Transactions on Affective Computing*, (2020)

[4] **Escalante, H.J., H. Kaya, A. Salah, S. Escalera, Y. Gucluturk, U. Guclu, X. Baró, I. Guyon, J. Junior, M. Madadi, S. Ayache, E. Viegas, F. Gürpınar, A. Wicaksana, C. Liem, M. Gerven and R. Lier,** Explaining first impressions: Modeling, recognizing, and explaining apparent personality from videos, *IEEE Transactions on Affective Computing*, (2018), 1–18.

[5] **Eyben, F., M. Wöllmer and B. Schuller,** Opensmile: the munich versatile and fast open-source audio feature extractor, *Proceedings of the 18th ACM international conference on Multimedia*, (2010), 1459–1462.

[6] **Eyben, F., K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan and K. Truong,** The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing, *IEEE Transactions on Affective Computing*, **7-2** (2015), 190–202.

[7] **Han, J., Z. Zhang, G. Keren and B. Schuller,** Emotion recognition in speech with latent discriminative representations learning, *Acta Acustica united with Acustica*, **104-5**, (2018), 737–740.

[8] **Han, J., Z. Zhang, Z. Ren and B. W. Schuller,** EmoBed: Strengthening monomodal emotion recognition via training with crossmodal emotion embeddings, *IEEE Transactions on Affective Computing*, (2019).

[9] **Hoffer, E. and N. Ailon,** Deep metric learning using triplet network, *International Workshop on Similarity-Based Pattern Recognition*, (2015), 84–92.

[10] **Jacques Junior, J.C.S., Y. Güçlütürk, M. Pérez, U. Güçlü, C. Andujar, X. Baró, H. J. Escalante, I. Guyon, M. A. J. Van Gerven, R. Van Lier and others,** First impressions: A survey on vision-based apparent personality trait analysis, *IEEE Transactions on Affective Computing*, (2019), 1–20.

[11] **Kang, C., S. Xiang, S. Liao, C. Xu and C. Pan,** Learning consistent feature representation for cross-modal multimedia retrieval, *IEEE Transactions on Multimedia*, **17-3** (2015), 370–381.

[12] **Kampman, O., E.J. Barezi, D. Bertero and P. Fung,** Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (2018), 606—611.

[13] **Kaya, H., F. Gurpinar and A.A. Salah,** Multimodal score fusion and decision trees for explainable automatic job candidate screening from video CVS, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, (2017), 1–9.

[14] **Kingma, D.P. and J. Ba,** Adam: A method for stochastic optimization, arXiv preprint (2014), https://arxiv.org/pdf/1412.6980.pdf

[15] **Li, Y., J. Wan, Q. Miao, S. Escalera, H. Fang, H. Chen, X. Qi and G. Guo,** CR-Net: A deep classification-regression network for multimodal apparent personality analysis, *International Journal of Computer Vision*, (2020), 1–18.

[16] **Ngiam, J., A. Khosla, M. Kim, J. Nam, H. Lee and A.Y. Ng,** Multimodal deep learning, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, (2011), 689–696.

[17] **Ponce-López, V., B. Chen, M. Oliu, C. Corneanu, A. Clapés, I. Guyon, X. Baró, H. J. Escalante and S. Escalera,** ChaLearn LAP 2016: First round challenge on first impressions - Dataset and results, *Computer Vision – ECCV 2016 Workshops, Lecture Notes in Computer Science*, (2016), 400–418.

[18] **Raffel, C. and D.P.W. Ellis,** Feed-forward networks with attention can solve some long-term memory problems, arXiv preprint (2016), https://arxiv.org/pdf/1512.08756.pdf

[19] **Tsai, Y.-H.H., P.P. Liang, A. Zadeh, L.-P. Morency and R. Salakhutdinov,** Learning Factorized multimodal representations, *International Conference on Learning Representations (ICLR)*, (2019), 1–20.

[20] **Wang, B., Y. Yang, X. Xu, A. Hanjalic and H.T. Shen,** Adversarial cross-modal retrieval, *Proceedings of the 25th ACM international conference on Multimedia*, (2017), 154–162.

[21] **Wang, X., X. Han, W. Huang, D. Dong and M.R. Scott,** Multi-Similarity loss with general pair weighting for deep metric learning, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2019), 5022–5030.

[22] **Willis, J. and A. Todorov,** First impressions making up your mind after a 100-ms exposure to a face, *Psychological Science*, **17-7** (2006), 592-598.

[23] **Wimmer, M., B. Schuller, D. Arsic, G. Rigoll and B. Radig,** Low-level fusion of audio, video feature for multi-modal emotion recognition, *Proceedings of the Third International Conference on Computer Vision Theory and Applications (VISAPP)*, (2008), 145—151.

[24] **Wöllmer, M., A. Metallinou, F. Eyben, B. Schuller and S.S. Narayanan,** Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling, *11th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, (2010), 1–4.

[25] **Xintao, W., Y. Ke, D. Chao and L.C. Chen,** Recovering realistic texture in image super-resolution by deep spatial feature transform, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018)

[26] **Zadeh, A., M. Chen, S. Poria, E. Cambria and L.-P. Morency,** Tensor fusion network for multimodal sentiment analysis, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, (2017), 1103–1114.

[27] **Zeng, Z., M. Pantic, G.I. Roisman and T.S. Huang,** A survey of affect recognition methods: Audio, visual, and spontaneous expressions, *IEEE transactions on pattern analysis and machine intelligence*, **31-1**, (2008), 39–58.

[28] **Zhang, L., S. Peng and S. Winkler,** PersEmoN: A deep network for joint analysis of apparent personality, emotion and their relationship, *IEEE Transactions on Affective Computing*, (2019), 1–10.

Á. Fodor, A. Lőrincz and R.R. Saboundji
Department of Artificial Intelligence
Faculty of Informatics
Eötvös Loránd University
Budapest
Hungary
foauaai@inf.elte.hu
lorincz@inf.elte.hu
sxdj3m@inf.elte.hu