Multimodal Sentiment and Personality Perception Under Speech: A Comparison of Transformer-based Architectures

Ádám Fodor FOAUAAI@INF.ELTE.HU
Eötvös Loránd University, Hungary

Rachid R. Saboundji SXDJ3M@INF.ELTE.HU
Eötvös Loránd University, Hungary

Julio C. S. Jacques Junior JJACQUES@CVC.UAB.CAT
Computer Vision Center, Spain

Sergio Escalera SERGIO@MAIA.UB.ES
Computer Vision Center & University of Barcelona, Spain

David Gallardo DAVID.GALLARDO@UB.EDU
University of Barcelona, Spain

Andras Lőrincz LORINCZ@INF.ELTE.HU
Eötvös Loránd University, Hungary

Abstract

Human-machine, human-robot interaction, and collaboration appear in diverse fields, from homecare to Cyber-Physical Systems. Technological development is fast, whereas real-time methods for social communication analysis that can measure small changes in sentiment and personality states, including visual, acoustic and language modalities are lagging, particularly when the goal is to build robust, appearance invariant, and fair methods. We study and compare methods capable of fusing modalities while satisfying real-time and invariant appearance conditions. We compare state-of-the-art transformer architectures in sentiment estimation and introduce them in the much less explored field of personality perception. We show that the architectures perform differently on automatic sentiment and personality perception, suggesting that each task may be better captured/modeled by a particular method. Our work calls attention to the attractive properties of the linear versions of the transformer architectures. In particular, we show that the best results are achieved by fusing the different architectures' preprocessing methods. However, they pose extreme conditions in computation power and energy consumption for real-time computations for quadratic transformers due to their memory requirements. In turn, linear transformers pave the way for quantifying small changes in sentiment estimation and personality perception for real-time social communications for machines and robots.

Keywords: Linear Transformers; Multimodal information fusion; Personality perception; Sentiment analysis; Fairness.

1. Introduction

In social communication, emotion and personality are two of the main cues that define how we feel and behave when we interact with others, as well as how we (and others) react while interacting. If we aim intelligent systems to engage and communicate with us in a "human-like" way, we need to provide them with similar mechanisms. This may explain why automatic sentiment, emotion and personality perception (and recognition) are receiving increasing interest from the Machine Learning (ML) and Computer Vision (CV) communities (Vinciarelli and Mohammadi, 2014; Jacques Junior et al., 2019; Zeng et al., 2009; Soleymani et al., 2017; Avots et al., 2019), but also Personality Psychology (PP) (Phan and Rauthmann, 2021), which is supported by the easy access to large and public databases and benchmarks on related topics, in addition to the outstanding advancements of deep learning methods.

The research on human affective behavior and personality computing has evolved from single modality (e.g., audio, visual, or text-based) to multimodal, including audiovisual fusion, linguistic and paralinguistic fusion, and multicue visual fusion based on facial expressions, head movements, body gestures (Zeng et al., 2009) or context (Martinez, 2019). However, past publicly available datasets and works on these fields, especially those focusing on visual cues, were not explicitly addressing the "under speech" scenario, where facial expressions produced during speech could lead to low recognition performances on different tasks if the dynamics of the expression are not considered. That is, most works and public datasets found in the literature for visual emotion recognition are focused on still images, e.g., (Guo et al., 2018). At the same time, information gathered "under speech" time intervals is most important during interactions and can express a rich variety of intentions (Hellbernd and Sammler, 2016). To address this problem, recent works and datasets started to pay more attention to temporal dynamics such as the CMU-MOSEI dataset, developed for the case of multimodal sentiment analysis and emotion recognition (Zadeh et al., 2018).

Previous works used to model temporal dynamics in these kinds of applications through frame-by-fame recognition followed by aggregation (Biel et al., 2012; Celiktutan and Gunes, 2014). Nonetheless, such approach may not generalize well under speech due to the dynamics of the task. Moreover, human face-to-face communication is a complex multimodal signal, where words (language modality), gestures/pose/gaze (visual modality) and changes in tone (acoustic modality) are used to convey our intentions (Zadeh et al., 2018). To address this problem, different deep learning-based solutions have been proposed, from Recurrent Neural Networks (Majumder et al., 2019; Elbarougy et al., 2020; Schoneveld et al., 2021) to the recently emerging Transformers (Vaswani et al., 2017), as proposed by Palmero et al. (2021) or Siriwardhana et al. (2020). Past works have already exploited the use of Transformer architecture for personality recognition (Leonardi et al., 2020), but in the context of Natural Language Provessing (NLP). In this work, we follow the latter multimodal Transformer-based approaches and evaluate different Transformer architectures on state-of-the-art databases for the problem of sentiment and personality perception under speech, with the purpose of finding the trade-off between accuracy and efficiency.

According to Avots et al. (2019), most of the existing approaches for emotion recognition are tailored for a specific database, which also applies for personality computing and sentiment analysis. This way, while the model is trained on a particular database, it usually

faces high variation in appearance and background, as well as human-centered attributes, such as head pose, ethnicity, age, gender, and local fashion, which in turn impose a strong limitation when the goal is to build robust and fair machine learning models which are capable of generalizing well to different populations, contexts and environments. Although erasing demographic bias in data-driven machine learning architectures is of extreme difficulty, in this work we encode visual information as Action Units (or AUs, for short) (Ekman and Rosenberg, 1997; Baltrušaitis et al., 2015) to mitigate such problem. It must be emphasized that using AUs as features are neither supposed nor intended to remove all kinds of bias. Our goal here is to be sure that the model is at least not basing its decisions on some appearance-based features like the background, people's skin tone or hair stile. Action Unit detectors might have their own biases, depending on the way they were trained. Nevertheless, we believe such kind of feature can also promote cross-dataset generalization.

Beyond the desire for competitive performance, applications *must be real-time* while satisfying *privacy-related constraints* in many environments giving rise to the combined demand of fast processing and low memory consumption. Cloud-based computation is possible under the condition that information is de-identified *before* it is sent to the cloud. Evaluations in the cloud using Action Unit estimations seem plausible at first sight, but facial dynamics may still reveal identity (Stone, 2001), limiting computations to local computers. Cloud service for speech recognition is more complex. Even if the speech is transformed to a robotic one (Fodor et al., 2021) before being sent, personal information concerning names, places, and dates set serious barriers. These points highlight the need for local computations and the potential advantages of the linear version of the transformer family.

In this work we are focusing on real-time applications. This choice restricts us in the selection of the tools. For example, we are not using features developed for specific databases not available otherwise. In turn, although competitive, our results could be improved further and are sometimes surpassed by others. Our contributions are summarized as follows:

- We compare different Transformer-based architectures, with the aim of finding the trade-off between accuracy and efficiency, as well as the best competitive architecture that works well for both sentiment and personality perception;
- We use Action Units as inputs for the visual modality, which are supposed to be invariant to some appearance-based features (e.g., skin-tone or hair-style). Thus, mitigating possible sources of bias toward under-represented groups/categories while promoting cross-dataset/domain/scenario generalization.
- We consider memory and speed parameters as required by real-time processing in human-machine, human-robot interactions, an important goal of future developments.
- The proposed approach was evaluated on different tasks and datasets. The results obtained are similar to those given by state-of-the-art models on the respective tasks and datasets.

We note that the underlying machine learning technology (i.e., the technology of information fusion using deep neural networks) has diverse applications and is developing quickly, leading to more elaborated and efficient architectures and performance improvements. We show that real-time applications are feasible today even if resources concerning GPU strength and memory consumption are limited.

The paper is organized as follows. In Sec. 2, we present a brief overview about the state-of-the-art on sentiment and personality perception with respect to methodologies and datasets, with a particular focus on visual and deep learning-based approaches under speech. The evaluated architectures are presented in Sec. 3. Sections 4 and 5 respectively describe and discuss the experiments and the results. Finally, we draw the conclusion and make our final remarks in Sec. 6.

2. Related Work

2.1. Sentiment

The ML and CV communities have widely studied the analysis and understanding of people's affective state over the past years (Zeng et al., 2009; Soleymani et al., 2017). The majority of works found in the literature and publicly available datasets on the topic deal with instantaneous expression categorization (Guo et al., 2018), where the task is to classify a discrete affective state using features from different modalities (Soleymani et al., 2017). From the past few years, however, the research community started to pay attention to the design and development of novel multimodal datasets and annotation protocols to advance state of the art on the field, taking into account how people communicate and express ideas and opinions through verbal content as well as visual and vocal features, such as facial expressions, head gestures, and voice quality (Zadeh et al., 2016a).

In the work of Zadeh et al. (2016a), the CMU Multimodal Opinion-Level Sentiment Intensity (CMU-MOSI) dataset was introduced, a video corpus with opinion-level sentiment intensity annotations that can be used for sentiment, subjectivity, and multimodal language studies. On the basis of interaction patterns between words and facial gestures, the authors presented a simple representation model that jointly accounts for words and gestures in each opinion segment. As a baseline, they trained prediction models using Support Vector Regression, showing that the proposed multimodal dictionary can yield better results in sentiment intensity analysis compared with common fusion methods. Later, Zadeh et al. (2018) introduced the CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset, which is currently the largest dataset of multimodal sentiment analysis and emotion recognition. They also presented a Multi-attention Recurrent Network (MARN) model for understanding human communication. According to these authors, the main strength of MARN comes from discovering interactions between modalities over time using a Multi-attention Block and storing them in the hybrid memory of a recurrent component called the Long-short Term Hybrid Memory, which is an extension of the Long-short Term Memory by reformulating the memory component to carry hybrid information. Similarly to our work, Action Units are used as indicators of facial muscles movements. However, a set of visual features including per-frame basic and advanced emotions are also considered in their work.

Siriwardhana et al. (2020) represented text, acoustic (speech), and visual modalities with features extracted independently from pre-trained Self Supervised Learning (SSL) models, applied to multimodal emotion recognition. Given the high dimensional nature of SSL-based features, they introduced a novel Transformer and Attention-based fusion mechanism to efficiently combine and train on multimodal embeddings. They achieved state-of-the-art results for sentiment and emotion recognition on different datasets. For the visual modality,

they used a pre-trained FAb-Net (Wiles et al., 2018) model to obtain embeddings for each frame in the video that contained the speaker's face. Their work was partially inspired by the Multimodal Transformer (MulT) using unaligned language sentences (Tsai et al., 2019). To the best of our knowledge, the current state-of-the-art on sentiment recognition on CMU-MOSI/MOSEI is the work of Han et al. (2021). In their work, a Bi-Bimodal Fusion Network (BBFN) is proposed, an end-to-end network that performs fusion (relevance increment) and separation (difference increment) on pairwise modality representations. For the visual modality, MulT and BBFN use the *FACET* module of the commercial iMotion software, and make the extracted features available for research purposes. The FACET module estimates 35 AUs. Thus, this also mitigates possible sources of bias coming from (raw or deep) appearance-based features.

Our literature review on sentiment perception revealed that state-or-the-art methods on the topic are considering the use of visual inputs based on Action Units (Zadeh et al., 2018; Han et al., 2021), that could alleviate possible sources of bias coming from appearance-based features. This is not the typical route in automatic personality perception since features like hairstyle, piercings, and makeups are excellent tools for self-expression. Furthermore, to the best of our knowledge, we are the first to evaluate an efficient linear transformer for the task of audio-visual sentiment perception.

2.2. Personality

Personality computing (Vinciarelli and Mohammadi, 2014) is receiving high attention from different research communities due to its applicability on emerging Human-Centered Artificial Intelligence scenarios. It covers automatic personality recognition, perception, and synthesis, which may guide the machine or the robot in interactive, collaborative scenarios.

There is almost no difference in how the ML and CV communities are addressing automatic personality recognition or perception when it comes to supervised learning. That is, the main difference between them is the origin of the labels. In the case of recognition (Palmero et al., 2021), the labels are generally obtained from self-report questionnaires. In the case of perception (Jacques Junior et al., 2019), the labels are given from the perspective of external observers, usually obtained via crowd-sourcing or other-report questionnaires. Both cases currently share a common limitation, i.e., large and publicly available databases on these topics are scarce, constraining advancing the field's state of the art.

The ChaLearn First Impression Challenge, designed with the purpose of advancing the research on the field, introduced one of the largest publicly available datasets on the topic of personality perception, i.e., the First Impressions (Ponce-Lopez et al., 2016) dataset. It is composed of 10K short-video clips (around 15s long) of people talking to the camera, annotated with Big-Five (apparent) personality traits. It was further extended (Escalante et al., 2017) with the inclusion of transcripts, gender, ethnicity annotations, and an additional "invite to interview" variable. Although different kinds of bias have been found (Escalante et al., 2020; Jacques Junior et al., 2021) in the First Impressions dataset, from the classical unbalanced distribution bias problem with respect to different attributes to perception (subjective) biases with respect to gender, age, ethnicity and face attractiveness coming from crowdsource based annotations, it is still being broadly used to advance the research

on this field. It should be emphasized that being able to identify, understand and explain those biases can be part of the problem.

Multimodal personality perception has been addressed and studied by the computational research community in different ways. In the work of Gürpınar et al. (2016), a pre-trained Convolutional Neural Network (CNN) was employed for extracting facial expressions as well as scene information. Visual features representing facial expressions and scenes were combined and fed to a Kernel Extreme Learning Machine (ELM) regressor. The work has been extended in (Kaya et al., 2017) to consider audiovisual information. Principi et al. (2019) studied different sources of biases affecting personality perception also using the First Impressions dataset, including emotions from facial expressions, attractiveness, age, gender, and ethnicity, as well as their influence on prediction ability of apparent personality. Li et al. (2020) presented a deep Classification-Regression Network (CR-Net) based on visual, acoustic, and textual information. In their work, both the entire scene and the face of the person are analyzed using ResNet-34 (He et al., 2015) as the backbone. The authors also introduced a Bell Loss function to address inaccurate predictions caused by the regression-to-the-mean problem.

More recently, Palmero et al. (2021) introduced UDIVA, a non-acted dataset consisting of 90.5 hours of face-to-face dyadic interactions, where both self-reported and apparent personality labels are given. As a baseline, they proposed a transformer-based method for regressing the self-reported personality traits of a target person, taking into account multimodal information from both participants in the interaction obtained from 3-sec video segments. The work was extended in (Curto et al., 2021) by considering variable time windows, which allowed the modeling of longer-term interdependencies, and a cross-subject layer, which enables the network to explicitly model interactions through attentional operations. They also proposed to model the behavior of both individuals simultaneously, through a two-stream cross-attentional Transformer, to predict their personalities jointly eventually.

Different from the above works, where raw visual information or deep features obtained from raw data are given as part of the input to the network to regress the (apparent or self-reported) personality of a subject, we use Action Units (AUs) as "visual" input, being more generic and thus allaying possible sources of bias coming from appearance-based features. Although AU-based features are not a standard when it comes to personality recognition or perception, Wu et al. (2020) have already proposed to exploit it. In addition of using acoustic and textual information, they proposed to extract a set of visual features including facial action units, facial landmarks, head pose, gaze tracking and HOG features, proposing a many2many interaction scheme to perform time-dependent interactions explicitly inside single modality and across different modalities. Similarly, we adopt AUs as one of the main sources for visual information. Furthermore, and as far as we know, there is no similar study evaluating an efficient linear transformer for the task of multimodal personality perception.

3. Methods

First, we review the key components of the multimodal Transformer architecture. Then, we compare the quadratic and linear attention modules that will be evaluated in this work.

Finally, we briefly describe the method that extends the architecture to self-supervised pre-processing networks.

Each task is trained/evaluated individually on different datasets, as detailed in Sec. 4. Each evaluated architecture has been adapted to have a final layer responsible for regressing either the sentiment score or the personality trait values. The latter, in a multi-task fashion or trait-wise, depending on the training strategy (detailed in Sec. 4.3).

Next, we highlight the underlying main concepts followed by relevant details of the evaluated architectures.

3.1. Transformer architecture: the key components

The essential element of a cross-modal transformer is called single-head transformer (Fig. 1), which builds upon the attention model made of (i) queries of the modality to be transformed to, and (ii) keys and values of the modality being the source of the transformation. Single-head transformer is repeated many times, giving rise to a multi-head system (Fig. 2(a)). Multi-head unit receives normalized embedded input modulated by a genuine positional encoding method (Vaswani et al., 2017). Multi-head outputs are combined with their inputs via skip connections followed by additions, normalization, and a position-wise feedforward network, crucial for extracting temporal information from different time instants.

The key thought that simplifies the quadratic transformer (Tsai et al., 2019) to its linear version is described next. Additional details about the linear version of the quadratic transformer, beyond the ones presented here, can be found in (Katharopoulos et al., 2020).

3.2. Comparison of the Quadratic and the Linear Transformers

In our work, we consider three data modalities: acoustic, visual and textual components. For each target modality, we can generate two cross-modal transformers (using the other two inputs as sources), which outputs are combined by concatenation before going through a self-attention transformer (Fig. 2(b)). Self-attention highlights and weights the related elements, giving a combined representation. This procedure is repeated for each modality. Then, the outputs of the three self-attention modules are combined before a fully-connected layer with linear activation producing the final prediction of the multimodal transformer.

3.2.1. Multimodal Transformer (MulT)

The Multimodal (Quadratic) Transformer (Tsai et al., 2019) is an end-to-end model that extends the standard Transformer network (Vaswani et al., 2017) to a multimodal setting. The model is built up from multiple stacks of pairwise and bidirectional attention blocks, that effectively implement the fusion process.

In the following, we compare the quadratic attention network of MulT with the linear version (Lin-MulT) of attention (Katharopoulos et al., 2020). First, the input sequences are passed through a 1D temporal convolution layer to preserve the local structure:

$$\hat{X}_{(\cdot)} = \text{Conv1D}(X_{(\cdot)}, k_{(\cdot)}) \in \mathbb{R}^{l_{(\cdot)} \times d}, \tag{1}$$

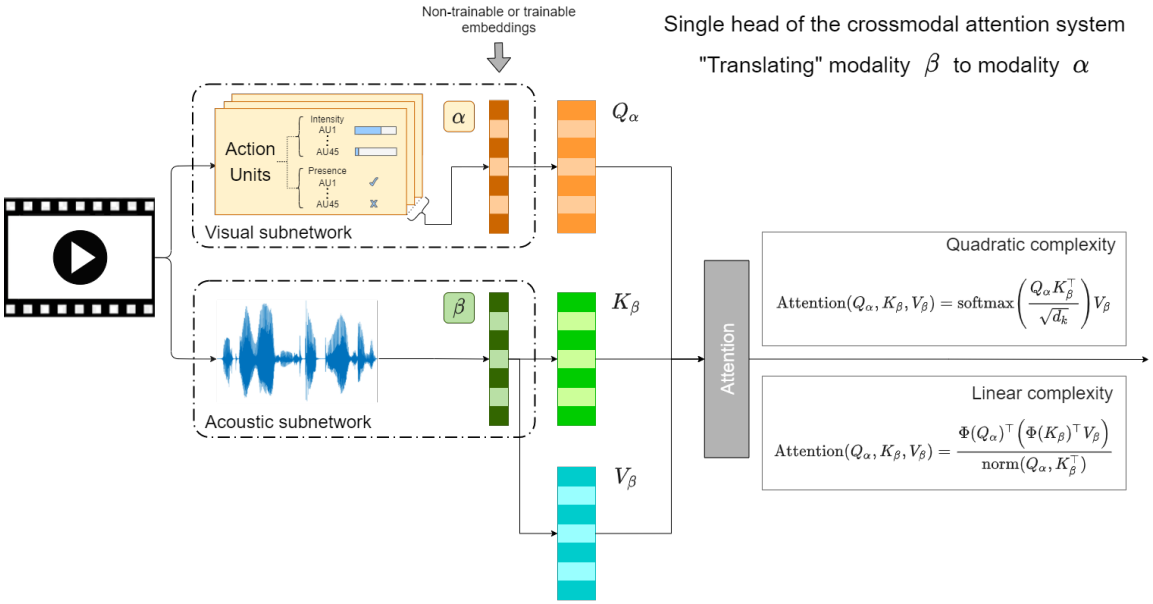

Figure 1: Single head of a transformer. Transformers translate one information source (e.g., β, here associated with the acoustic modality) to another one (e.g., α, here associated with the visual modality). Embeddings indicated by darker striped columns can be features derived from the raw data or the outputs of a pre-trained deep model. Transformers learn keys (K_β) and values (V_β) of modality β and the queries (Q_α) from the α modality. These three quantities denoted by lighter striped columns form the core of the attention system of a single head. Attention is computed differently in linear and quadratic transformers, sketched in the boxes on the right.

where $k_{(\cdot)}$ are kernel sizes of the convolutional layers and $(\cdot)$ stands for acoustic (A), visual (V) or textual (T) modality. Positional encoding is added for implicitly encoding temporal information (Vaswani et al., 2017).

For cross-modal attention translating modality β to modality α, inputs are transformed to α-queries, β-keys, and β-values by means of trainable matrices as follows

$$
\begin{aligned}
Q_{(\alpha)} &= \hat{X}_{(\alpha)} W_{Q_{(\alpha)}}, \\
K_{(\beta)} &= \hat{X}_{(\beta)} W_{K_{(\beta)}}, \\
V_{(\beta)} &= \hat{X}_{(\beta)} W_{V_{(\beta)}}, \\
V'_{(\alpha)} &= \text{softmax}\left(\frac{Q_{(\alpha)} K_{(\beta)}^\top}{\sqrt{d_k}}\right) V_{(\beta)}.
\end{aligned}
\tag{2}
$$

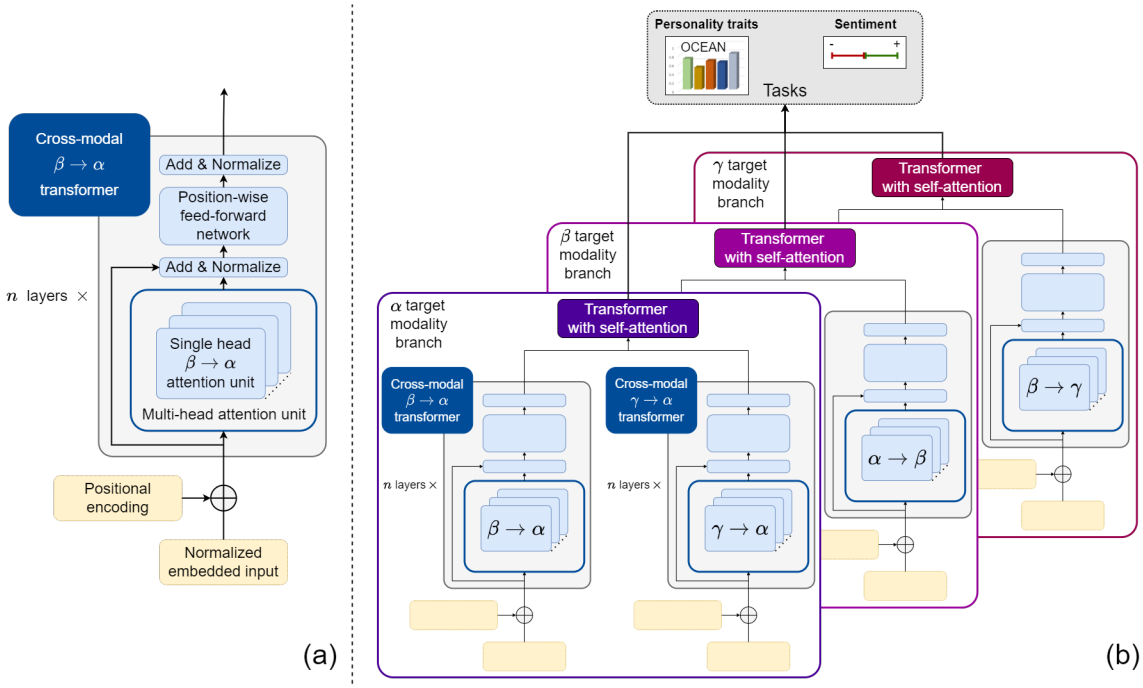

Figure 2: (a) Standard Multi-head attention unit. (b) Multimodal transformer: source modalities i and j are transformed to target modality k. Such two units are combined by another transformer network that utilizes self-attention to fuse the information pieces to form a branch within the multimodal network before outputting the predicted score(s).

3.2.2. Multimodal Transformer with Linear Attention (Lin-MulT)

Multimodal Transformer with Linear Attention reformulates and generalizes Eqs. 2 by introducing the notation $\text{sim}(q, k)$, the similarity function between the query and the key as the exponentiation of the dot product:

$$\text{sim}(q, k) = \exp\left(\frac{q^\top k}{\sqrt{d_k}}\right). \tag{3}$$

Given that subscripting a matrix with i returns the i^{th} row as a vector, Katharopoulos et al. (2020) generalized attention equation for any similarity function as follows:

$$V'_{(\alpha)_i} = \frac{\sum_{j=1}^{l_{(\beta)}} \text{sim}\left(Q_{(\alpha)_i}, K_{(\beta)_j}\right) V^\top_{(\beta)_j}}{\sum_{j=1}^{l_{(\beta)}} \text{sim}\left(Q_{(\alpha)_i}, K_{(\beta)_j}\right)}. \tag{4}$$

The only constraint for the "sim" function is its non-negativity. Kernel $k(x, y) : \mathbb{R}^{2\times d} \rightarrow \mathbb{R}_+$ satisfies this constraint and writing $k(x, y) = (\Phi(x), \Phi(y))$ one can rewrite Eq. 4 as

$$V'_{(\alpha)_i} = \frac{\Phi(Q_{(\alpha)_i})^\top \sum_{j=1}^{l_{(\beta)}} \Phi(K_{(\beta)_j}) V_{(\beta)_j}^\top}{\Phi(Q_{(\alpha)_i})^\top \sum_{j=1}^{l_{(\beta)}} \Phi(K_{(\beta)_j})} \tag{5}$$

Considerable speed gain arises since one may compute $\sum_{j=1}^{l_{(\beta)}} \Phi(K_{(\beta)_j}) V_{(\beta)_j}^\top$ only once and reuse these quantities in every query. We followed (Katharopoulos et al., 2020) and applied $\Phi(x) = \text{elu}(x) + 1$ in the computations.

3.2.3. Self Supervised Embedding Fusion Transformer (SSE-FT)

Self Supervised Embedding Fusion Transformer (Siriwardhana et al., 2020) uses a modified version of the fusion technique described in MulT. While the cross-modal multi-head attention unit is the same, the differences are (i) in the application of independently pre-trained Self-Supervised Learning models for feature extraction (ii) in the self-attention method applied for two of the inputs, speech and video, before the multi-head attention module, (iii) in the compressed trainable extension of the SSE called CLS (classification) token, being part of the input and used as Query instead of the full sequence, and (iv) Hadamard product is used for CLS token fusion before the final prediction.

The first token of each sequence is the CLS token, which can aggregate the information embedded in the entire sequence. The idea of CLS token is also motivated by NLP literature; it is used by RoBERTa (Liu et al., 2019) to represent an entire sequence. Therefore, only acoustic and visual embeddings are prepended by this unique token.

The cross-modal attention module called IMA is basically the same as described in Eq. 2, with one difference: the Query (Q) vector are created from the CLS token of one modality, while Key (K) and Value (V) vectors are computed from the entire sequence of the other modality within a pair.

The modality pairs of cross-modal units within a branch are the same as in MulT and Lin-MulT. Source modalities (i, j) are transformed to a target modality (k): $i \rightarrow k$ and $j \rightarrow k$. It is repeated for all the pairs (i.e., audio-visual, audio-text, and visual-text).

The CLS token can be used as a compressed representation to solve downstream tasks. In the case of the three input modalities, the six CLS tokens of the three branches are used as inputs to a straightforward late fusion mechanism. However, the CLS tokens are not just concatenated, as mutual information is extracted more efficiently among pairs of CLS representations (using the same target modality) with the Hadamard product. The resulting representations are concatenated and sent through the prediction layer.

We are including SSE-FT method in our comparison because of the potential benefits of SSE features (described in Sec. 4.3), the use of its end-to-end architecture in future works and most importantly, the added value of its facial features information.

4. Experimental Setup

In this section, we describe the feature extraction methods (Sec. 4.1), the datasets and evaluation protocol used for each task (Sec. 4.2), and the training strategy (Sec. 4.3).

4.1. Data Preprocessing

4.1.1. Acoustic Features

eGeMAPS: extended Geneva Minimalistic Acoustic Parameter Set (Eyben et al., 2015) contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, Hammarberg index, and slope V0 features. Only low-level audio descriptors are used, considering voiced and unvoiced regions. The audio signals are extracted from the videos using FFmpeg with 44100 sampling frequency. Then, eGeMAPS LLDs are generated using OpenSMILE (Eyben et al., 2010), resulting in 25 features for every audio frame. Standardization is applied as a preprocessing step.

Wav2Vec: A multi-layer CNN called Wav2Vec (Schneider et al., 2019) is used as deep audio feature extractor. It was trained on 960 hours of audio taken from the Librispeech (Panayotov et al., 2015) dataset. Self-supervised training leverages the concept of Contrastive Predictive coding. The 512-dimensional context representation of Wav2Vec efficiently represents the raw audio waveform[1].

4.1.2. Visual Features

AUs: Action Units are used to avoid introducing unnecessary sources of bias to the model by excluding the background and other appearance-based information that could be retrieved from raw visual data such as gender, age, or ethnicity. Action Units (AU) (Baltrušaitis et al., 2015) construct facial expressions encoded in the Facial Action Code System (FACS). These AUs can be described in two ways: presence (indicating whether a particular AU is detected in a given time frame) and intensity (indicating how intense an AU is at a given time frame). For this purpose, we used OpenFace (Baltrušaitis et al., 2018), an open-source toolkit. We extracted 35 Action Unit presence and intensity values per frame, which are standardized before further usage. A possible alternative to OpenFace is FACET (used in Han et al. (2021)), however, FACET module from the commercial iMotion software is not publicly available for research purposes.

FAb-Net: Facial Attributes-Net (Wiles et al., 2018), a visual deep feature extractor, trained on VoxCeleb1 and VoxCeleb2 video datasets[2]. The network learns embedding that encodes facial attributes like landmarks, poses, and emotions without any labels in a self-supervised manner. It is used on all the frames of a given video (30 fps) where the speaker's face is detected. The extracted embedding has a size of 256 for each frame.

4.1.3. Textual Features

BERT: Bidirectional Encoder Representations from Transformers (Devlin et al., 2018) is a powerful transformer architecture with self-attention that learns contextual relations between words (or sub-words) in a text. It is bidirectional, which exploits the context from both left and right to extract patterns or representations during training. BERT is applied for extracting high-level representations from textual data. We use the HuggingFace[3] im-

1. Pre-trained weights are available at https://github.com/pytorch/fairseq
2. The model with the pre-trained weights are available at https://github.com/oawiles/FAb-Net
3. https://github.com/huggingface

plementation, which computes 768-dimensional context-dependent word embeddings from the transcripts.

RoBERTa: Robustly optimized BERT approach (RoBERTa) (Liu et al., 2019) is pre-trained on raw texts using a set of large English-language datasets without human annotations.

RoBERTa is available using the Fairseq sequence modeling toolkit[1]. As tokenization, Byte-Pair Encoding (BPE) is applied to the transcript, then a 1024-dimensional feature vector represents context-aware semantic information of each word.

4.1.4. Self-Supervised Embeddings (SSE)

Self-Supervised Embeddings are features extracted from pre-trained Self-Supervised Learning (SSL) models. In our experiments, three frozen SSL models, namely Wav2Vec (Schneider et al., 2019), FAb-Net (Wiles et al., 2018), and RoBERTa (Liu et al., 2019) are used for extracting the acoustic, visual, and textual SSE embeddings, respectively. The models' pre-trained weights are publicly available, no further finetuning is applied. The generated input representations are then combined in different ways, as detailed in Sec. 4.3.

4.2. Datasets and Evaluation Protocol

The architectures detailed in Sec. 3 have been trained and evaluated on different tasks and databases, as detailed next.

CMU-MOSI: Multimodal Opinion-level Sentiment Intensity (MOSI) dataset (Zadeh et al., 2016b) contains 93 videos with a total of 2199 utterances. Each utterance has a continuous sentiment intensity label in the range of $[-3, +3]$. The original split was used in the experiments: 52, 10, and 31 videos in the training, validation, and test sets, with 1151, 296, and 752 utterances, respectively.

CMU-MOSEI: Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) (Amir et al., 2018) is one of the largest datasets for multimodal emotion and sentiment recognition. CMU-MOSEI consists of over 23,500 utterances created by extracting review videos from YouTube. Each utterance is annotated for the sentiment with a continuous score in the range of $[-3, +3]$. The samples are also annotated with seven emotion class labels. However, we only consider sentiment annotation in this work. We followed the original train, valid, test split available in CMU-Multimodal-SDK[4].

First Impression: For personality perception, we used the ChaLearn First Impressions (FI) database (Ponce-Lopez et al., 2016), which is the largest publicly available in-the-wild dataset on the topic. The FI dataset was released in the context of a computational challenge, where the goal was to automatically recognise the Big-Five (OCEAN) apparent personality traits of single individuals in videos: ***O****penness to experience*, ***C****onscientiousness*, ***E****xtraversion*, ***A****greeableness*, and ***N****euroticism*[5]. The dataset comprises 10K short video

4. https://github.com/A2Zadeh/CMU-MultimodalSDK
5. *Neuroticism* was labelled in Ponce-Lopez et al. (2016) as *"Emotion stability"*, which is the opposite of Neuroticism. This will be represented later in Sec. 5 as $\overline{\text{N}}$.

clips (average duration of 15s each, divided into training, validation, and test sets) extracted from more than 3K different YouTube videos of people talking to a camera. It was annotated using crowdsourcing through pairwise comparisons, which were further converted to continuous values (Chen et al., 2016). That is, the adopted method individually converts the ordinal ratings of each dimension into continuous values (such as the level of "*Extraversion*") by fitting a Bradley-Terry-Luce (BTL) model with maximum likelihood, which is further scaled to be in the range of $[0, 1]$. This way, each video sample in the dataset will have a continuous value associated with each trait dimension, which can be used by any supervised learning method in a classification or regression task. In a nutshell, the labels correspond to human annotators' impressions. Therefore, the Big-five traits are considered as first impressions or perceived personalities. This is a different task than evaluating self-reported personality traits (as in Palmero et al. (2021)), yet it is just as important in human interactions. Later, the FI dataset was extended (Escalante et al., 2017) with the inclusion of transcripts in addition to an "invite to interview" variable, gender, and ethnicity information, aiming to advance research on explainable machine learning. The original training/validation/test split has been used for our experiments.

Evaluation Metrics: Following the state-of-the-art works on each task and dataset, we use the following metrics:

- **CMU-MOSI/CMU-MOSEI:** we use 7-class accuracy, 2-class accuracy (binary), Mean Average Error (MAE), F1-score, and Pearson correlation between the prediction and target;
- **First Impressions:** "1-Mean Absolute Error" is the performance metric, and R^2 represents the coefficient of determination;

4.3. Training Strategy

Taking into account computational and memory restrictions, we use different sets of features (Table 1) for the experiments based on the adopted architecture, detailed in Sec. 5.

Set	Features
OOB	eGeMAPS, AUs, BERT
WFR	Wav2Vec, FAb-Net, RoBERTa
OO-WFR	eGeMAPS, AUs, Wav2Vec, FAb-Net, RoBERTa

Table 1: Multimodal feature sets: we shall use the "OOB" shorthand for the **O**penSMILE (for the eGeMAPS features), **O**penFace (for the Action Units based features), and **B**ERT feature combination, and "WFR" for the **W**av2Vec, **F**Ab-Net, and **R**oBERTa feature combination. "OO-WFR" is used as the combination of handcrafted and deep features (introduced in Sec. 4.1).

For CMU-MOSI and CMU-MOSEI, we set a maximum sequence length of 10 seconds for audio-visual modalities. For the ChaLearn FI dataset, we consider 15 seconds as sequence length. Sequences shorter than 10 (or 15) seconds are padded with zeros and longer ones are truncated. We trained Lin-MulT in a multi-task (MT) and trait-wise (TW) manner to measure the performance gain of either procedure in combating the regression-to-the-mean

problem, which is unique to the FI dataset dataset, taking into account the adopted ones. While multi-task learning benefits from shared early representations in general, this claim does not necessarily hold for the First Impressions dataset.

MulT and Lin-MulT: For the input sequences, we use 1D convolution to capture temporal information and reduce dimensionality for further computations. We chose 32 as the number of kernels and 8 the number of attention heads. We consistently used five stacked transformer layers with dropout (0.2 chance) throughout the experiments related to different tasks. The only architectural difference between MulT and Linear-MulT is the attention type: MulT uses the quadratic, while Linear-MulT applies a linear attention mechanism (detailed in Sec. 3.2). Both models were trained using Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$) and a learning rate of $2e-3$. We used the Bell Loss (Li et al., 2020) to further combat the regression-to-the-mean problem. When the loss on the validation set plateaus, the learning rate is divided by a factor of 10.

SSE-FT: The full architecture is implemented from Fairseq code-base[6]. We used Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 1e-6$) with warm-up updates, learning rate of $3e-4$ and polynomial decay as a learning-rate scheduler. Transformer hyperparameters for CMU-MOSI/MOSEI are the same as in (Siriwardhana et al., 2020).

5. Experiments and Results

This section compares the results obtained by the different models described in Sec. 3 on different tasks and databases. The compared methods are the Multimodal Transformer (MulT), the Multimodal Transformer with Linear Attention (Lin-MulT) and the Self-Supervised Embedding Fusion Transformer (SSE-FT).

We present results on sentiment regression using the CMU-MOSI and CMU-MOSEI datasets in Sec. 5.1. We compare and discuss MulT *vs.* Lin-MulT on the same task, to demonstrate the competitiveness of the linear attention mechanism. In Sec. 5.2, we present and discuss obtained results on the personality perception task, using the First Impressions dataset. Linear transformer offers efficient fusion methods, and we exploit this property by combining OOB and WFR features (detailed in Sec. 4.3). This would be infeasible for the quadratic transformers, taking into account computational and memory requirements. The importance of individual target modality branches in cross-modal transformers is tested in Sec. 5.3. State-of-the-art results on different tasks and datasets are also reported and briefly discussed. Finally, we conclude the experimental section with a short discussion regarding real-time considerations, computational and memory constraints of the performed experiments in Sec. 5.4.

5.1. Sentiment Analysis

First, we compare Lin-MulT and MulT on both CMU-MOSI and CMU-MOSEI and demonstrate performance scaling with respect to dataset size. Then, we analyze and discuss the obtained results of all evaluated methods for the same task and datasets. Note that the methods here evaluated for sentiment estimation use either OOB or WFR features, detailed

6. https://github.com/shamanez/Self-Supervised-Embedding-Fusion-Transformer

in Sec. 4.3. The obtained results on CMU-MOSI and CMU-MOSEI are shown in Table 2 and Table 3, respectively.

Method (Feature)	Acc7 ↑	Acc2 ↑	F1 ↑	MAE ↓	Corr ↑
MulT (OOB)	0.382	0.803	0.803	0.889	0.597
Lin-MulT (OOB)	0.354	0.784	**0.864**	0.956	0.679
SSE-FT (WFR)	**0.463**	**0.838**	0.835	**0.76**	**0.75**
MARN (Zadeh et al., 2018)	0.347	0.771	0.770	0.968	0.625
MulT (Tsai et al., 2019)	0.401	0.833	0.829	0.832	0.745
SSE-FT (Siriwardhana et al., 2020)	**0.465**	0.839	0.835	**0.776**	**0.768**
BBFN (Han et al., 2021)	0.450	**0.843**	**0.843**	**0.776**	0.755

Table 2: Results for multimodal sentiment analysis on CMU-MOSI. The rows on top: evaluated methods. The rows at the bottom: state-of-the-art results obtained from the literature. Notations: ↓ (↑) shows that lower (higher) values are better. Best results (per metric) for both the evaluated methods and those found in the literature are highlighted in bold.

Method (Feature)	Acc7 ↑	Acc2 ↑	F1 ↑	MAE ↓	Corr ↑
MulT (OOB)	0.416	0.671	0.671	0.76	0.240
Lin-MulT (OOB)	0.417	0.741	0.743	0.778	0.476
SSE-FT (WFR)	**0.540**	**0.850**	**0.820**	**0.580**	**0.570**
MulT (Tsai et al., 2019)	0.511	0.845	0.845	0.570	0.758
SSE-FT (Siriwardhana et al., 2020)	**0.557**	**0.873**	**0.870**	**0.529**	**0.792**
BBFN (Han et al., 2021)†	0.548	0.862	0.861	**0.529**	0.767

Table 3: Results for multimodal sentiment analysis on CMU-MOSEI. The rows on top: evaluated methods. The rows at the bottom: state-of-the-art results obtained from the literature. Notations: ↓ (↑) shows that lower (higher) values are better. Best results (per metric) for both the evaluated methods and those found in the literature are highlighted in bold.

In the case of MOSI (Table 2), Lin-MulT achieved better performance on F1 and Corr metrics, while being slightly worse on Acc7 and Acc2. MulT outperformed Lin-MulT on MAE by a high margin. We hypothesize that Lin-MulT achieved better precision/recall trade-off between all classes, illustrated by a higher F1/Corr score. This, in turn, is less favorable if accuracy metrics are preferred.

In the case of MOSEI (Table 3), results follow a similar trend when MulT and Lin-MulT are considered. That is, MulT performed better than Lin-MulT on MAE metric. Nevertheless, an improvement of at least 6% in Acc2, F1 and the Corr metrics can be observed in favor of Lin-MulT. These results highlight the benefits of linear attention in terms of performance scalability with respect to the number of samples in a dataset, taking into account that CMU-MOSEI is a much larger database, as well as in the the precision/recall trade-off.

Performance of SSE-FT was closely reproduced on CMU-MOSI (Table 2), compared with the original method (Siriwardhana et al., 2020). Although, this was not the case for CMU-MOSEI (Table 3). Hyperparameter tuning may explain the difficulty of reproducing their results. Overall, SSE-FT model obtained the best results on CMU-MOSI and CMU-

MOSEI, taking into account the evaluated architectures (and features). The only exception was the case of the F1 score (on CMU-MOSI), where Lin-MulT obtained slightly better performance. One possible reason for SSE-FT high performance is the use of Self-Supervised Learned features, since MulT and Lin-MulT rely on handcrafted ones. However, it should be noticed that SSE-FT model relies on the quadratic attention network, being computationally and memory expensive.

In comparison with state of the art, MulT and Lin-MulT obtained overall better results than MARN (Zadeh et al., 2018), which also uses AU as visual input (i.e., appearance-invariant features). However, neither MulT nor Lin-MulT obtained better results than BBFN (Han et al., 2021) on CMU-MOSI (Table 2) and CMU-MOSEI (Table 3). Nevertheless, BBFN method uses a bi-bimodal fusion consisting of two (quadratic) Transformer-based bimodal learning modules, being also computationally/memory expensive.

5.2. Personality Perception

Table 4 shows the results for personality perception on the ChaLearn First Impressions (Ponce-Lopez et al., 2016) dataset, obtained from distinct architectures and feature sets. We also analysed two different training strategies, i.e., multi-task (MT) and trait by trait training, or simply trait-wise (TW) for short.

First, it should be noted that all evaluated methods presented a similar performance on the FI dataset, as shown in Table 4. This is a well known problem related with the First Impressions dataset, where most personality scores are centered in a very small region close to the mean. 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 accuracy at test stage (Escalante et al., 2020) due to the highly centralized distribution. In turn, changes in the third digit are relevant.

To ease the analysis, we discuss the results of Table 4 in a pairwise manner, given the method (MulT, Lin-MulT and SSE-FT), feature set (OOB, WFR and OO-WFR) and training strategy (MT or TW).

- **OOB *vs.* WFR:** surprisingly, the worse results in the case of personality perception were obtained using the WFR features, specifically when using Lin-Mult and SSE-FT architectures, both trained in a multi-task fashion. Following this tendency and the same training strategy, Lin-MulT (OOB) obtained overall better results than Lin-Mult (WFR). These results were not observed for the case of sentiment analysis, previously discussed. Contrary, SSE-FT with WFR features obtained the best results for sentiment estimation. This suggests that different tasks may be better modeled/captured by different feature sets and model architectures.

- **MulT *vs.* Lin-MulT:** in this scenario, MulT performed overall slightly better than Lin-MulT, using the same set of features (OOB) and training strategy (MT). Nevertheless, we consider Lin-MulT still competitive as it saves computational and memory resources, motivating us to exploit different feature combinations such as OO-WFR (discussed next). Note that feature fusion like OO-WFR could be prohibitive when the quadratic attention module is combined with limited computational resources.

- **MT *vs.* TW:** before discussing OO-WFR feature fusion, we evaluated training some of the models for personality perception trait-wise, instead of multi-task. We observed that the trait-wise helped to improve the results, for instance, if we compare Lin-MulT using OOB feature set on both cases. This is reinforced when we combine different features as mentioned above (OO-WFR) using the same method. That is, Lin-MulT using OO-WFR set obtained even better performance when trained trait by trait, compared to multi-task training, also suggesting that both feature sets (OOB and WFR) bring complementary information. Nevertheless, this improvement is not typical as multi-task training usually benefits training with complementary information from the different tasks (or traits, in our case). We hypothesize this may be related with the fact that the FI dataset has the scores centered in a very small region close to the mean.

Method (Feature)	Strategy	O ↑	C ↑	E ↑	A ↑	$\overline{\text{N}}$ ↑	Avg ↑	R^2 ↑
Lin-MulT (WFR)	MT	0.9007	0.8967	0.8966	0.9019	0.8941	0.8980	0.24
SSE-FT (WFR)	MT	0.9022	0.8954	0.8972	0.9046	0.8982	0.8995	0.26
Lin-MulT (OOB)	MT	0.9041	0.8960	0.9037	0.9005	0.8988	0.9006	0.27
MulT (OOB)	MT	0.9033	0.8975	0.9036	0.9024	0.9010	0.9016	0.29
Lin-MulT (OOB)	TW	0.9052	0.8979	0.9045	0.9031	0.9008	0.9023	0.30
Lin-MulT (OO-WFR)	MT	0.9074	0.9004	0.9073	0.9043	0.9021	0.9043	0.32
Lin-MulT (OO-WFR)	TW	**0.9078**	**0.9007**	**0.9077**	**0.9058**	**0.9041**	**0.9050**	**0.33**
CR-Net(†) (Li et al., 2020)	MT	**0.9195**	**0.9218**	**0.9202**	**0.9177**	**0.9146**	**0.9188**	-

Table 4: Personality perception results on the First Impressions (Ponce-Lopez et al., 2016) dataset: ↑ shows that higher values are better. Best results for both the evaluated methods and those found in the literature are highlighted in bold.

In comparison with the state of the art, CR-Net (Li et al., 2020) still gives better results compared to the evaluated architectures, feature sets and training strategies. However, CR-Net works under a particular and different condition we are trying to avoid, i.e., it uses raw visual data (from the entire scene and the person's face) that can be sensitive to appearance-based features. In contrast, the evaluated approaches use Action Units and FAb-Net (Wiles et al., 2018) features as visual inputs that do not capture identity-specific information according to the literature (Agarwal et al., 2020). Moreover, the evaluated approaches/features may promote cross-dataset/domain/scenario generalization.

5.3. Cross-modal ablation

We measured the importance of individual target modality branches in cross-modal transformers on MulT and Lin-MulT using the ChaLearn First Impressions dataset. According to Table 5 and R^2 metric, cross-modal transformer with video as target modality had the best results for MulT, whereas the one with audio as the target was the winner for Lin-MulT. We found no specific target modality whose cross-modal transformer outperformed the rest.

Source → Target	Method	Avg (1-MAE) ↑	R2 ↑
A+V → T	MulT	**0.902**	**0.151**
	Lin-MulT	0.894	0.143
T+A → V	MulT	**0.900**	**0.244**
	Lin-MulT	0.899	0.221
V+T → A	MulT	0.900	0.239
	Lin-MulT	**0.901**	**0.246**

Table 5: Cross-modal ablation on the First Impressions (Ponce-Lopez et al., 2016) dataset, given a source and target modality.

5.4. Computational and memory constraints

Real-time evaluation constrains memory consumption. Linear transformers scale better than quadratic ones, and drastic changes may rise in computational depending on the available GPU units since most computations can be parallel. Nevertheless, energy consumption will scale linearly and quadratically in the two methods, a considerable difference in some IoT applications.

Beyond the scaling properties, exact numbers can change quickly due to the quick technological advances. In turn, our description is restricted. We used an RTX Titan unit with 24 GB RAM. Computation of a 10-second sample took 48 ms and 52 ms for the linear and the quadratic transformers, respectively. For a 30-second (100-second) sample, we had 51 ms (61) for the linear and 150 ms (1400 ms) for the quadratic one, respectively. The drastic change in time means that the quadratic method ran out of parallel computational units. Beyond these computations, preprocessing also limits the number of available GPU units.

Progress in technology may offer solutions. For example, FPGA accelerators are about 7 times to 45 times faster than comparable systems using high-end GPUs, and the FPGAs may consume ten times less energy for deep CNN networks, like YOLOv3 (Hesse, 2021; Wei et al., 2021). Tensor Processing Units (TPUs) can alleviate constraints on real-time computations. In addition, cloud computing can also mitigate the problem, although privacy issues may arise. In all of these cases, the problem of energy consumption remains.

To close this discussion, it should be mentioned that our RTX Titan unit with 24 GB RAM was insufficient to test the MulT (quadratic) method for more than three channels to be fused; we quickly ran out of memory. The advancements of deep technology will give rise to more sophisticated and efficient transformer-based fusion methods. Nonetheless, the linear advantage will hold.

6. Final Remarks

Transformer models are state-of-the-art in deep learning technology; they are overcoming Convolutional Neural Networks (CNNs) on different tasks (Dosovitskiy et al., 2020). We compared the fusion capabilities of state-of-the-art transformer models, the Transformer-Based Self Supervised Feature Fusion Model (SSE-FT) (Siriwardhana et al., 2020) and the Multimodal Transformer Model (MulT) (Tsai et al., 2019) on sentiment estimation and personality perception databases. We also compared the performance of quadratic attention of these models and the case when we replaced these attention models with the linear version in MulT. We restricted our studies to methods that are appearance invariant, with

the purpose of mitigating possible sources of bias coming from appearance-based features. Our main findings are summarized as follows:

- On sentiment estimation, MulT, and SSE-FT performed similarly, but with the increase of the database, SSE-FT performed better;
- On personality perception, where the distribution of the data is very narrow, MulT was better than SSE-FT;
- The linear attention version of MulT (Lin-MulT) was competitive in all cases;
- Lin-MulT is computationally efficient. The linear scaling of its memory requirement, instead of the quadratic one, enabled the fusion of the features that gave better results for the case of personality perception;

All in all, memory, speed, accuracy, and fusing capabilities of the Lin-MulT architecture make it an attractive choice for real-time sentiment analysis and personality perception.

Acknowledgments

This work has been partially supported by EU H2020 project Humane AI Net (grant agreement No. 952026), by Spanish project PID2019-105093GB-I00 and by ICREA under the ICREA Academia programme.

References

Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2020.

Zadeh Amir, Liang Paul Pu, Poria Soujanya, Cambria Erik, and Morency Louis-Philippe. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 2236–2246, 2018.

Egils Avots, Tomasz Sapiński, Maie Bachmann, and Dorota Kamińska. Audiovisual emotion recognition in wild. *Machine Vision and Applications*, 30(5):975–985, Jul 2019.

Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic Action Unit detection. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 06, pages 1–6, 2015.

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face and Gesture Recognition*, 2018.

Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. FaceTube: predicting personality from facial expressions of emotion in online conversational video. In *ICMI*, pages 53–56, 2012.

Oya Celiktutan and Hatice Gunes. Continuous prediction of perceived traits and social dimensions in space and time. In *ICIP*, pages 4196–4200, 2014.

Baiyu Chen, Sergio Escalera, Isabelle Guyon, Víctor Ponce-López, Nihar Shah, and Marc Oliu. Overcoming calibration problems in pattern labeling with pairwise ratings: Application to personality traits. In *European Conference on Computer Vision Workshop (ECCVW)*, pages 419–432, 2016.

David Curto, Albert Clapés, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, David Gallardo-Pujol, Georgina Guilera, David Leiva, Thomas B. Moeslund, Sergio Escalera, and Cristina Palmero. Dyadformer: A multi-modal transformer for long-range modeling of dyadic interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2177–2188, October 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

Reda Elbarougy, Bagus Tris Atmaja, and Masato Akagi. Continuous audiovisual emotion recognition using feature selection and LSTM. *Journal of Signal Processing*, 24(6):229–235, 2020.

Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio C. S. Jacques Junior, Meysam Madadi, Xavier Baró, Stephane Ayache, Evelyne Viegas, Yağmur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven, and Rob van Lier. Design of an explainable machine learning challenge for video interviews. In *International Joint Conference on Neural Networks (IJCNN)*, pages 3688–3695, 2017.

Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Julio Cezar Silveira Jacques Junior, Meysam Madadi, Xavier Baró, Stéphane Ayache, Evelyne Viegas, Yagmur Güçlütürk, Umut Güçlü, Marcel A. J. van Gerven, and Rob van Lier. Design of an explainable machine learning challenge for video interviews. In *International Joint Conference on Neural Networks (IJCNN)*, pages 3688–3695, 2017.

Hugo Jair Escalante, Heysem Kaya, Albert Salah, Sergio Escalera, Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, Isabelle Guyon, Julio C. S. Jacques Junior, Meysam Madadi, Stephane Ayache, Evelyne Viegas, Furkan Gurpinar, Achmadnoer S. Wicaksana, Cynthia Liem, Marcel A. J. Van Gerven, and Rob Van Lier. Modeling, recognizing, and explaining apparent personality from videos. *IEEE Transactions on Affective Computing*, 2020.

Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenSMILE: The Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, 2010.

Florian Eyben, Klaus Scherer, Bjorn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202, 2015.

Ádám Fodor, László Kopácsi, Zoltán Ádám Milacski, and András Lőrincz. Speech de-identification with deep neural networks. *Acta Cybernetica*, 25(2):257–269, 2021.

Jianzhu Guo, Zhen Lei, Jun Wan, Egils Avots, Noushin Hajarolasvadi, Boris Knyazev, Artem Kuharenko, Julio C. Silveira Jacques Junior, Xavier Baró, Hasan Demirel, Sergio Escalera, Jüri Allik, and Gholamreza Anbarjafari. Dominant and complementary emotion recognition from still images of faces. *IEEE Access*, 6:26391–26403, 2018.

Furkan Gürpınar, Heysem Kaya, and Albert Ali Salah. Combining deep facial and ambient features for first impression estimation. In Gang Hua and Hervé Jégou, editors, *ECCVW*, pages 372–385, 2016.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, page 6–15, New York, NY, USA, 2021. Association for Computing Machinery.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

Nele Hellbernd and Daniela Sammler. Prosody conveys speaker's intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, 88:70–86, 2016.

Christopher Noel Hesse. Analysis and comparison of performance and power consumption of neural networks on CPU, GPU, TPU and FPGA, 2021.

Julio C. S. Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A. J. van Gerven, Rob van Lier, and Sergio Escalera. First impressions: A survey on vision-based apparent personality trait analysis. *IEEE Transactions on Affective Computing*, 2019.

Julio C. S. Jacques Junior, Agata Lapedriza, Cristina Palmero, Xavier Baro, and Sergio Escalera. Person perception biases exposed: Revisiting the First Impressions dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 13–21, January 2021.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

Heysem Kaya, Furkan Gürpinar, and Albert A. Salah. Multi-modal score fusion and decision trees for explainable automatic job candidate screening from video CVs. In *CVPRW*, pages 1651–1659, 2017.

Simone Leonardi, Diego Monti, Giuseppe Rizzo, and Maurizio Morisio. Multilingual transformer-based personality traits estimation. *Information*, 11(4), 2020.

Yunan Li, Jun Wan, Qiguang Miao, Sergio Escalera, Huijuan Fang, Huizhou Chen, Xiangda Qi, and Guodong Guo. CR-Net: A deep classification-regression network for multimodal apparent personality analysis. *International Journal of Computer Vision*, 128(12):2763–2780, Dec 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 6818–6825, Jul. 2019.

Aleix Martinez. Context may reveal how you feel. *Proceedings of the National Academy of Sciences*, 116:201902661, 03 2019.

Cristina Palmero, Javier Selva, Sorina Smeureanu, Julio C. S. Jacques Junior, Albert Clapes, Alexa Mosegui, Zejian Zhang, David Gallardo, Georgina Guilera, David Leiva, and Sergio Escalera. Context-aware personality inference in dyadic scenarios: Introducing the UDIVA dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 1–12, January 2021.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

Le Vy Phan and John Rauthmann. Personality computing: New frontiers in personality assessment. *Social and Personality Psychology Compass*, 15, 06 2021.

Victor Ponce-Lopez, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, and Sergio Escalera. ChaLearn LAP 2016: First round challenge on First Impressions - dataset and results. In *European Conference on Computer Vision (ECCV) Workshop*, pages 400–418, 2016.

Ricardo Darío Pérez Principi, Cristina Palmero, Julio C Junior, and Sergio Escalera. On the effect of observed subject biases in apparent personality analysis from audio-visual signals. *IEEE Transactions on Affective Computing*, 2019.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Liam Schoneveld, Alice Othmani, and Hazem Abdelkawy. Leveraging recent advances in deep learning for audio-visual emotion recognition. *Pattern Recognition Letters*, 146:1–7, 2021.

Shamane Siriwardhana, Tharindu Kaluarachchi, Mark Billinghurst, and Suranga Nanayakkara. Multimodal emotion recognition with transformer-based self supervised feature fusion. *IEEE Access*, 8:176274–176285, 2020.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

Jim Stone. Face recognition: When a nod is better than a wink. *Current Biology*, 11(16): R663–R664, 2001.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Transaction on Affective Computing*, 5(3):273–291, 2014.

Kaijie Wei, Koki Honda, and Hideharu Amano. An implementation methodology for neural network on a low-end FPGA board. *International Journal of Networking and Computing*, 11(2):172–197, 2021.

Olivia Wiles, A Koepke, and Andrew Zisserman. Self-supervised learning of a facial attribute embedding from video. *arXiv preprint arXiv:1808.06882*, 2018.

Liangqing Wu, Dong Zhang, Qiyuan Liu, Shoushan Li, and Guodong Zhou. Speaker personality recognition with multimodal explicit many2many interactions. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2020.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016a.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *ArXiv*, abs/1606.06259, 2016b.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Zhihong Zeng, Maja Pantic, Glenn I. Roisman, and Thomas S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.