ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Multimodal Recognition of Valence, Arousal and Dominance via Late-Fusion of Text, Audio and Facial Expressions

Annette Rios[1], Uwe Reichel[2], Chirag Bhuvaneshwara[3], Panagiotis Filntisis[4], Petros Maragos[4], Felix Burkhardt[2], Florian Eyben[2], Björn Schuller[2], Fabrizio Nunnari[3] and Sarah Ebling[1] *

1 - University of Zurich
Department of Computational Linguistics - Switzerland

2 - audEERING GmbH - Germany

3 - German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus, D3.2 - Germany

4 - ATHENA Research Center - Greece

Abstract. We present an approach for the prediction of valence, arousal, and dominance of people communicating via text/audio/video streams for a translation from and to sign languages. The approach consists of the fusion of the output of three CNN-based models dedicated to the analysis of text, audio, and facial expressions. Our experiments show that any combination of two or three modalities increases prediction performance for valence and arousal.

1 Introduction

Affect Detection or Emotion Recognition has received attention in research on AI with a variety of application scenarios in mind, from emotion-aware chatbots to analysis of consumer behaviour [1]; it is the task of guessing information about the emotional state of a person based on their utterances, appearance, and behavior. For an application to sign language translation, our goal is to extract affective information from speaking/writing people to modulate the motion style and facial expression of sign language avatars; thus producing an output that is more expressive and human-like. Taken alone, text, audio, and video modalities provide very limited results, but improvements are possible when data are fused with each other [2] in order to combine the respective strength of each modality. As an example, valence recognition has been shown good performance based on video and text features, whereas arousal can be better predicted from speech acoustics [3]. Fusion paradigms can be amongst others subdivided into *early* feature-level and *late* decision-level fusion [4]. In early fusion a multimodal model is fed with a concatenation of features from several modalities. In late fusion, the input to the multimodal model is given by the answers of several unimodal models. The latter approach allows for a more flexible integration of

*This work is funded by the EU Horizon 2020 programme within the EASIER project (Grant agreement ID: 101016982). https://www.project-easier.eu/

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

modules developed e.g. by different project partners and is therefore pursued in this study.

We target the development of a mobile app for sign language (SL) translation where users can provide input via either text, audio (speech), or video (speech or signing), and an avatar will animate the translated content. Since the final app used to record the message that need to be translated allows multiple modalities for input, we extract information about the emotional state of the user from audio, video or text, or any combination thereof. If more than one modality is available, the results are fused for a final prediction. Our experiments show that emotions are generally better predicted if more than one input source is available. Even though the project is multilingual, covering multiple signed and spoken languages, the emotion recognition so far only covers German.

Emotions are usually modelled either as categorical classes with different granularity (e.g., Ekman's universal emotions [5]) or as continuous values along a given set of dimensions (e.g., VAD, valence-arousal-dominance [6]). In this work, we predict on the VAD space by fusing the VAD output of the audio model together with the Ekman classes from the text and video models.

2 Emotion modelling

Speech We apply a fine-tuned wav2vec 2.0 transformer model developed by Wagner et al. [7]. It is a multitask regression model that outputs valence, arousal, and dominance scores between 0 and 1. The underlying pretrained model is *facebook/Wav2Vec2-Large-Robust* [8]. We fine-tune this model on the MSP-Podcast[1] corpus. As suggested by Wang et al. [9], the initial feature extractor was kept frozen and only the encoder weights were fine-tuned with a mean concordance correlation coefficient (CCC) loss over the three emotional dimensions. The model is available online.[2]

Text Since the targeted language of the text-based models is German, we transfer two data sets from English: MELD Friends [10] and GoEmotions [11]. MELD Friends consists of subtitles with timestamps, which allows us to transfer the annotated labels to the German subtitles via automatic alignment. For GoEmotions, we use machine translation to obtain the German version of the annotated data. Note that both approaches introduce a certain amount of noise. Both data sets are annotated with categorical labels.

We fine-tune pretrained language models from Huggingface [12] with a classifier head for all experiments. Any language model that can be loaded with the AutoModel interface and provides the option to use a classification head can be used in our setup.[3] We train all models with cross-entropy loss and early stopping on microF1[4] to avoid overfitting. Additionally, we use Layerwise Learning

[1] `https://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Podcast.html`

[2] `https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim`

[3] Code and transferred data available at `https://github.com/a-rios/emotion_detection`

[4] We use microF1 instead of macroF1 for early stopping since the classes in both datasets are heavily skewed in terms of frequency and we therefore expect microF1 to be a more accurate estimate of performance.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Rate Decay (LLRD) [13] for fine-tuning, which gives a small improvement (about $+1.0$ $F1_{micro}$) over a uniform learning rate (Decay rate 0.95, initial learning rate 2e-5.)

Video Our facial expression classification model is trained on the categorical, manual annotations from the AffectNet [14] database, which features a variety of people, camera and scene conditions. For our experiments, we include only the manually annotated images of the 7 Ekman emotions plus "Neutral" expression, resulting in a total of $287,654$ training images and $4,000$ validation set images. As the test set is not released, the validation set is split into two equal sets, which we use alternatively as validation and test.

We employ deep convolutional neural networks (CNNs) which are trained with stochastic gradient descent, using mini-batches of the input dataset. We experimented with different CNN architectures, including MobileNetV2 (which focuses on mobile application) [15] and EfficientNet-B0 [16]. All the models were pre-trained on Imagenet [17]) and fine-tuned on Affectnet [14]. The training images were normalized by cropping the face region, normalizing the color, and rotating around the center to bring the eyes at a horizontal level. Instead of using pre-trained face classification models, we configured our own training pipeline, aiming at matching state of the art results by optimizing image preprocessing schemes, rather than employing resource demanding architectures.

Fusion For the late fusion models, we use the German subset of the multimodal SEWA dataset [18] that was used, e.g., in the *AVEC 2017* multimodal affect recognition challenge [19]. This data set contains audio and video recordings with transcripts, i.e., it covers all three modalities used for the final emotion prediction. However, the SEWA dataset does not contain annotations for *dominance*. Therefore, specifically for this dimension, we additionally use the *Vera am Mittag* dataset (`https://sail.usc.edu/VAM/index.html`), in which video recordings are available for 5 speakers only. In order to make the prediction tasks uniform for both datasets, for the SEWA set we were using interpausal units as opposed to rolling windows of 6s length as in [19]. The ground truth segment-level emotion scores were obtained as suggested by [19] by averaging over the frame-based annotations within an optimized right-shifted window that compensates for annotation delays.

For each dimension (*valence*, *arousal*, and *dominance*, with values ranging from 0 to 1, we train 4 models: a K-Nearest neighbor regressor (KNN), a feed forward neural network (FF), a support vector regressor (SVR), and an X-treme Gradient Boosting Regressor (XGB). Additionally, we train a multitask FF for joint valence and arousal regression (FF-MT). The models are trained on the logits outputs of the unimodal text- and video-based models and the emotion scores of the audio-based model. Since the audio-based model was finetuned on English data, this fusion additionally serves as a cross-language transfer of emotion prediction to German. For the SEWA dataset (arousal and valence), we use the speaker-disjunct split of the AVEC 17 challenge. For the *Vera am Mittag* dataset (dominance), we create 3 speaker-disjunct partitions accordingly.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

We tune the hyperparameters of the KNN, SVR, and XGB models by Bayesian optimization in a 5-fold cross-validation on the merged training and development partition. The single- and multi-task FFs are set up with a batch normalization layer, a dense layer with a sigmoid activation function and a linear output projection. We set the dropout between the final hidden layer and the output projection to 0.1, and train the model with a batch size of 40 and a CCC loss. We keep the model with the highest CCC value on the development set. For the multitask model, we use the CCC mean over all emotion dimensions.

3 Results

Individual Modalities The validation of the audio-based emotion model is reported in detail in [7]. For space reasons we restrict the presentation of the unimodal evaluations to the text and video modality. The evaluation refers to the datasets the unimodal models were developed on.

As for the *text-based models*, we train on the transferred German data (see 2) with pretrained BERT models (`https://huggingface.co/dbmdz/bert-base-german-cased`). Performance on individual emotions mostly correlates with the frequency of the class in the training data, with some exceptions: For instance, models perform very well on *gratitude*, despite relatively low frequency in the training data. We suspect that *gratitude* might be relatively narrow in terms of lexical diversity. This makes the classification task somewhat easier for the model compared to a broad category such as *neutral*. Generally, the models trained on the transferred German data perform slightly worse than the models trained on the original English data (MicroF1 GoEmotions English: 55.4 vs. German: 52.6; MicroF1 MELD English: 64.6 vs. German: 55.1).

For the *video-based models*, the best image preprocessing configuration requires to first identify the face bounds, and then zoom out of a scaling factor 1.2 to include more of pixels surrounding immediately the face region. We reach 58.1% on the AffectNet dataset with the EfficientNet-B0 model. Considering that the EfficientNet model has fewer parameters [16], the performance of our best model is in-line with the state-of-the art performance on AffectNet dataset.

Fusion The late fusion validation results are reported in Table 1. In order to validate the benefit of a multimodal fusion approach compared to unimodal approaches, we additionally train and evaluate the four model types on unimodal outputs (i.e., emotion class logits and scores) as well as on all bimodal combinations as baseline models, which are also shown in Table 1. These results cannot be directly compared with the results reported in [19], since they are based on different segments as mentioned in section 2.

4 Discussion and conclusions

Table 1 clearly shows that multimodal late fusion approaches outperform the unimodal mappings of emotion class logits to emotion dimensions. For valence

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

	Valence			Arousal			Dominance		
	model	CCC	MAE	model	CCC	MAE	model	CCC	MAE
audio	FF-MT	.46	.10	FF-MT	.40	.09	SVR	.78	.07
text	FF-MT	.47	.11	FF-MT	.37	.10	SVR	.17	.13
video	FF-MT	.59	.10	FF-MT	.47	.10	SVR	.04	.11
audio-text	FF-MT	.52	.10	FF-MT	.44	.09	SVR	**.79**	**.07**
audio-video	FF-MT	.67	.08	FF-MT	.54	.09	XGB	.47	.10
text-video	FF-MT	.64	.10	FF-MT	.50	.10	KNN	.09	.14
all	FF-MT	**.70**	**.08**	FF-MT	**.57**	**.08**	XGB	.33	.11

Table 1: Performance in terms of CCC and MAE for emotion dimensions and for all modalities and their combinations. MAE can range from 0 to 1. For each modality (combination) the best performing model is presented.

and arousal, the trimodal approach performs best. For dominance, due to the sparse video training data, the bimodal audio plus text approach works best. The poor video-based dominance recognition might further be explained by its focus on facial expression, whereas according to [6] the perception of dominance is mainly influenced by a judgement of body and hand gestures. Furthermore, Table 1 shows that multitask regression of valence and arousal in late fusion works better than single task regressions, since all best models for these dimensions are multitask FeedForward neural networks.

The model, developed to be applied in bidirectional spoken-sign language translation, performs reasonably well also in video-only and video+text modes. However, it is not yet clear how would performances be when recognising emotions from faces performing sign language utterances, where the face is already full involved in the language production. This is part of our future investigations.

References

[1] Francisca Adoma Acheampong, Chen Wenyu, and Henry Nunoo-Mensah. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(7):e12189, 2020.

[2] Lisa Feldman Barrett et al. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychological Science in the Public Interest*, 20(1), July 2019.

[3] Björn W Schuller. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5), 2018.

[4] Zhihong Zeng et al. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(1), 2008.

[5] Paul Ekman. Facial expression and emotion. *American Psychologist*, 48(4):384–392, 1993.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

[6] Albert Mehrabian. *Basic Dimensions for a General Psychological Theory: Implications for Personality, Social, Environmental, and Developmental Studies*. Cambridge: Oelgeschlager, Gunn & Hain, 1980.

[7] Johannes Wagner et al. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *arXiv preprint arXiv:2203.07378*, 2022.

[8] Wei-Ning Hsu et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. *arXiv preprint arXiv:2104.01027*, 2021.

[9] Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding. *arXiv preprint arXiv:2111.02735*, 2021.

[10] Soujanya Poria et al. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 2019. ACL.

[11] Dorottya Demszky et al. GoEmotions: A dataset of fine-grained emotions. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 2020.

[12] Thomas Wolf et al. Transformers: State-of-the-art natural language processing. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online, 2020. ACL.

[13] Tianyi Zhang et al. Revisiting few-sample BERT fine-tuning. In *International Conference on Learning Representations*, 2021.

[14] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 10(1), 2019.

[15] Mark Sandler et al. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE conference on computer vision and pattern recognition*, 2018.

[16] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[17] Olga Russakovsky et al. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.

[18] Jean Kossaifi et al. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(3), 2019.

[19] Fabien Ringeval et al. Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proc. of the 7th annual workshop on audio/visual emotion challenge*, 2017.