MULTIMODAL END-TO-END GROUP EMOTION RECOGNITION USING CROSS-MODAL ATTENTION

Lev Evtodienko

Higher School of Economics

ABSTRACT

Classifying group-level emotions is a challenging task due to complexity of video, in which not only visual, but also audio information should be taken into consideration. Existing works on multimodal emotion recognition are using bulky approach, where pretrained neural networks are used as a feature extractors and then extracted features are being fused. However, this approach does not consider attributes of multimodal data and feature extractors cannot be fine-tuned for specific task which can be disadvantageous for overall model accuracy. To this end, our impact is twofold: (i) we train model end-to-end, which allows early layers of neural network to be adapted with taking into account later, fusion layers, of two modalities; (ii) all layers of our model was fine-tuned for downstream task of emotion recognition, so there were no need to train neural networks from scratch. Our model achieves best validation accuracy of 60.37% which is approximately 8.5% higher, than VGAF dataset baseline and is competitive with existing works, audio and video modalities.

CCS Concepts

•**Computing methodologies**∼**Artificial intelligence**∼ **Computer vision**; •**Human-centered computing**∼**Human computer interaction (HCI)**

1. INTRODUCTION

Emotion recognition is difficult and important task. Understanding emotions in groups of people is vital not only for every individual of a group, but also for people with different background and cultures. Moreover, knowledge about common emotion of group could be of interest for business, learning, healthcare, surveillance and robotics.

Early affective computing was focused on individuals, while in recent years more research is done for groups of people from raw, uncontrolled data (i.e. "in the wild") [1]. Results performed on these type of data are easier adoptable to real-world situations such as surveillance cameras or videos from internet.

Dataset used in training was presented in EmotiW 2020 Audio-Video Group Emotion Recognition grand challenge [2] and the exact challenge is Video level Group AFfect (VGAF). The task of this competition was to classify group

emotion into three different categories: negative, neutral and positive. The biggest challenges of this dataset are: different lightning conditions, languages, video quality, frame rate, occlusions and intersection of people. One of the approach to handle such data, is to use only one modality [3], [4]. Another option are two-stage models, where stages are feature extraction and modality fusion respectively [5].

Despite the fact, that unimodal models show decent results, when given poor input information, this problem can not be compensated by another modality, which will hurt the performance. As for two-stage models, fixed feature extractors can not be fine-tuned for a specific task with respect to the information from low-level fusion layers.

To address these issues we propose our model with following features:

- Model was trained fully end-to-end, compensating the problem of missing information about modalities interaction. Moreover, if one modality does not carry a lot of useful information, another one can mitigate this problem.

- We do not freeze layers throughout our model, which allows it to be fully optimized to a given task and helps model achieve solid results using models from different domains effectively.

2. RELATED WORK

Automatic human emotion recognition is a topic of active research for nearly a decade [6], [7], [8]. In early works on multimodal learning there were several techniques proposed, such as early fusion [9] and late fusion [10]. While the idea behind these techniques are simple, they show decent results and are still being widely adopted for multimodal tasks. Recently, researchers have been working not only on fusion techniques, but also on multimodal architectures, where pairs of different modalities' are being fed to network. These architectures, for instance, are Tensor Fusion Networks [11], LXMERT [12], ClipBERT [13], VATT [14], VILT [15]. Due to increase in computational power in recent years, a lot of work was done using multimodal learning in different areas, such as question
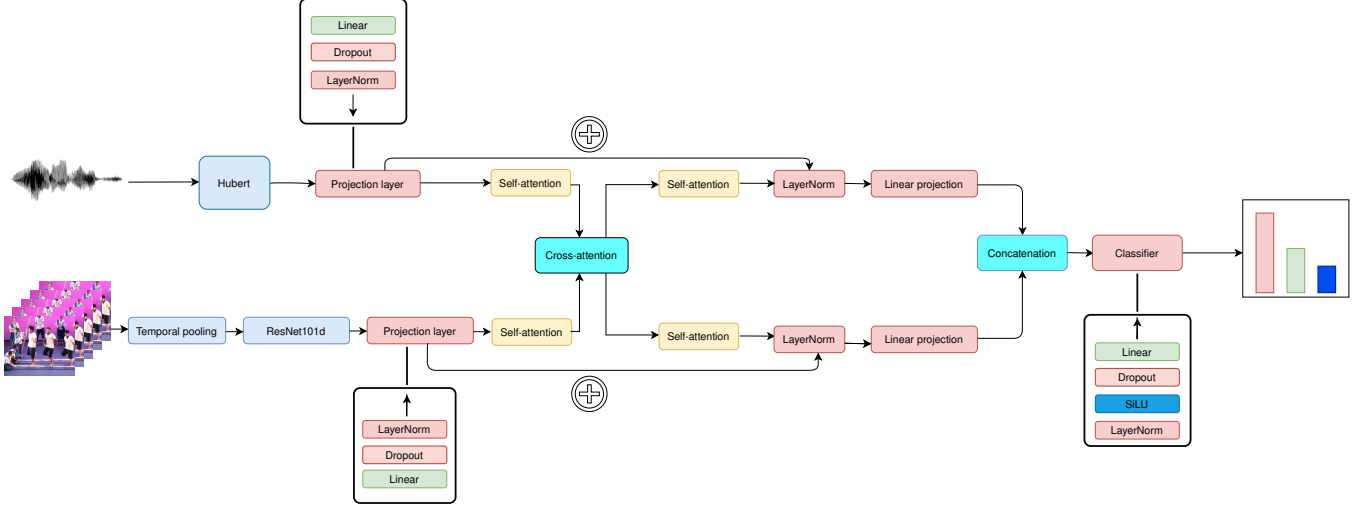
Fig. 1. Proposed end-to-end architecture for group emotion recognition.

answering [16], image captioning [17], emotion recognition [18], affect recognition [19].

Previous results on VGAF [2] dataset were obtained using two-stage models, where at first features were extracted using fixed models and then, late fusion was used for fusion of extracted features [5], [20], [21], [22], [23], [24], [25]. The best result for this dataset was shown by the winners of Audio-Visual Group Emotion Recognition challenge of EmotiW2020 [20], where team uses 5 different modalities and achieves 74.28% on validation set.

3. DATASET

The VGAF dataset [2] contains 2661 video clips for training and 766 for validation. Dataset was acquired from YouTube with tags, such as "fight", "holiday" or "protest", videos from which characterize different emotions. Each video was cropped in clip of 5 seconds length. The data is contains 3 classes – Positive, Neutral and Negative corresponding to the 3 emotion categories. Challenges of dataset are different contexts in every video, various resolutions, frame rates, multilingual audio, which is serious abstraction for the vast majority of available models. Moreover, as there are no labels of what exactly language is in the video, it makes it impossible to collect additional text transcription (modality) using automatic tools.

4. METHODOLOGY

4.1. Encoders

We describe the problem as follows. Let $X = \{(v_i, a_i)\}_{i=1}^I$ be a sample of data, where I is a number of multimodal samples, v_i is a sequence of RGB video frames and a_i is a raw audio of a given sample.

First, we extract 8 equally distributed frames. Vision encoder accepts a sequence of extracted frames $\forall v_i \in \mathbf{R}^{\mathbf{C} \times \mathbf{T} \times \mathbf{H} \times \mathbf{W}}$, where C, T, H, and W are the number of channels, temporal size, height, and width, respectively. For the vision encoder we use approach, inspired by ClipBERT [13], namely we applied mean-pooling (we will define this operation as M) to aggregate temporal information of a frames sequence, which is inexpensive way to make use of temporal information. We use pretrained 2D ResNet101d [26] for feature extraction. During our experiments we tried several backbone architectures described in **Table 1**. We decided not to use 3D CNNs, because it can highly increase time of training without advantage in accuracy [27], [28]. To pass extracted feature maps further, in attention layers, we flatten embedding on last two dimensions and proceed with passing vector to projection layer (defined as *PL* shown on Fig.1), which projected embeddings of different modalities to a common space. During current research instead of projection layer, layer module was considered, where GeLU [29] activation function and additional projection was added, but such a module lead to faster overfitting and shows approximately 3% drop in acccuracy.

Backbone	Accuracy
ResNet101d [26]	**60.37%**
ResNet50d [26]	58.44%
RexNet100 [30]	57.98%
ResNet50 [31]	56.83%

Table 1. Various vision encoder backbones and overall model validation accuracy.

We extracted raw audio data at sampling rate of 16000 Hz, and pass it to Hubert [32]. We choose this model, because it was trained in self-supervised manner, which provides more

robust representations and can handle multilingual data. To pass embeddings further, in self-attention layer, projection layer (shown on Fig.1) was used.

We define process of extracting embeddings of audio and video encoder stages as follows:

$$\forall v_i : v_{emb} = PL(ResNet101d(M(v_i)))$$
$$\forall a_i : a_{emb} = PL(Hubert(a_i))$$

for audio embeddings, where $v_{emb}, a_{emb} \in \mathbf{R^{S \times N}}$, S is a sequence length and N is a number of features.

4.2. Attention layers

In this section we will review attention layers used in our model: self-attention and cross-attention. The main purpose of these layers in our model is aligning multimodal information from encoding embedding after encoders.

Self-attention was initially intorduced in [33] and described as layer for extracting information from a set of *context* vectors to *query* vector. Formally, it can be written as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where K and V are context vectors and Q is a query vector. For our model we used multi-head attention (MHA), which was introduced in [34] and can be defined as

$$MHA(Q, K, V) = Concat(head_1, .., head_h)W^O$$

where $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$

W_i^Q, W_i^K, W_i^V are learnable parameters of query, key and value respectively. Unlike the BERT [35], in our model self-attention is not used for text data, but for audio and visual embeddings. In the context of our model input to the self-attention is $Q = K = V = (a_{emb} \oplus v_{emb})$.

Cross-modal attention has similar definition, but instead of computing dot product between the same vector for query and key, it makes use of multimodal nature of video, i.e. there are two cross-modal attentions, one takes K = V = a_{emb} and Q = v_{emb} as input; another one takes K = V = v_{emb} and Q = a_{emb}. Such attention enables one modality for receiving information from another modality and helps align them.

Following architecture decisions of [34], we have added **skip connections**, which add weights from layer before attention block with weights from layer after it. To prevent exploding gradient problem, Layer Normalization was applied before penultimate projection layers.

5. TRAINING

We train our model using Adam optimizer [36], with learning rate $1e^{-4}$ and weight decay $1e^{-3}$. As encoder parts of our models were already pretrained and only being fine-tuned
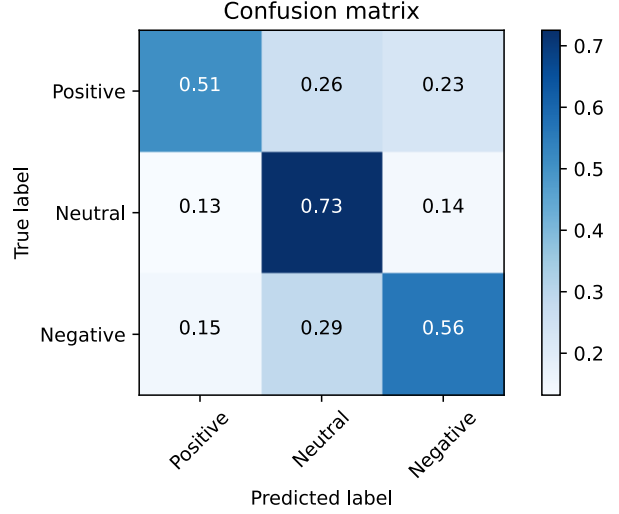

Fig. 2. Confusion matrix for predictions of our model.

during training procedure layers of this part of the model was trained with lower learning rate multiplied by factor of $1e^{-2}$.

One of the biggest challenge of learning multimodal neural networks is the fact, that they are exposed to severe overfitting. Usual regularization techniques are often ineffective [37] for these networks. To mitigate this problem we use **label smoothing** with $\epsilon = 0.2$, which makes neural network be less "confident" about class it predicts.

Models	Modality	Accuracy
K-injection network [21]	A+V	63.58%
ResNet50 + BiLSTM [25]	A+V	61.83%
Hubert+ResNet101d (ours)	A+V	**60.37%**
Inception + LSTM [2] (baseline)	A+V	52.09%

Table 2. Results on validation data for two modalities on VGAF dataset. A, V, F states for audio, video, face accordingly.

6. RESULTS

We use accuracy as evaluation metric of our model. It achieves an overall accuracy of 60.37% and outperforms baseline by a margin of 8.5%. In **Table 2** comparison of all available teams results, who used audio and video modalities for their final predictions, is shown. It can be seen, that our model has a competitive results compared to other works. The best model consists of ResNet101d and Hubert [32] as encoders for video and audio. The most challenging class for our model is "Positive", it can be explained by the challenging nature of the dataset and hard interpretation of emotions by themselves. Moreover, some videos in dataset have similar context, but different emotions and labels for them. Classifying "Neutral" videos as "Negative" ones can

be a problem, which emerges from the fact, that there are big number of protests in dataset, where some of them are peaceful and others are aggressive.

7. CONCLUSION

Group video emotion recognition is a challenging task, especially for "in the wild" data. In this paper we present model for VGAF dataset from Audio-Visual Group Emotion Recognition challenge of EmotiW2020. Two novel approaches is used for our model. Our model was trained end-to-end and optimized fully during training process, which help us achieve noticeable result of 60.37% validation accuracy, which outperforms baseline significantly and can perform practically on par with existing bimodal audio-visual models.

8. REFERENCES

[1] Abhinav Dhall, Roland Goecke, Shreya Ghosh, Jyoti Joshi, Jesse Hoey, and Tom Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, New York, NY, USA, 2017, ICMI '17, p. 524–528, Association for Computing Machinery.

[2] Abhinav Dhall, Garima Sharma, Roland Goecke, and Tom Gedeon, "Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 784–789, Association for Computing Machinery.

[3] Sandra Ottl, Shahin Amiriparian, Maurice Gerczuk, Vincent Karas, and Björn Schuller, "Group-level speech emotion recognition utilising deep spectrum features," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 821–826, Association for Computing Machinery.

[4] Andrey V Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," *arXiv preprint arXiv:2103.17107*, 2021.

[5] Boyang Tom Jin, Leila Abdelrahman, Cong Kevin Chen, and Amil Khanzada, "Fusical: Multimodal fusion for video sentiment," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 798–806, Association for Computing Machinery.

[6] Rosalind W. Picard, "Affective computing: From laughter to ieee," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 11–17, 2010.

[7] Panagiotis Tzirakis, George Trigeorgis, Mihalis A. Nicolaou, Björn W. Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *CoRR*, vol. abs/1704.08619, 2017.

[8] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "Iemocap: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[9] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, New York, NY, USA, 2011, ICMI '11, p. 169–176, Association for Computing Machinery.

[10] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency, "MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos," *CoRR*, vol. abs/1606.06259, 2016.

[11] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "Tensor fusion network for multimodal sentiment analysis," *CoRR*, vol. abs/1707.07250, 2017.

[12] Hao Tan and Mohit Bansal, "LXMERT: learning cross-modality encoder representations from transformers," *CoRR*, vol. abs/1908.07490, 2019.

[13] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu, "Less is more: Clipbert for video-and-language learning via sparse sampling," *CoRR*, vol. abs/2102.06183, 2021.

[14] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong, "VATT: transformers for multimodal self-supervised learning from raw video, audio and text," *CoRR*, vol. abs/2104.11178, 2021.

[15] Wonjae Kim, Bokyung Son, and Ildoo Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," 2021.

[16] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant, "Multimodalqa: Complex question answering over text, tables and images," *CoRR*, vol. abs/2104.06039, 2021.

[17] Iro Laina, Christian Rupprecht, and Nassir Navab, "Towards unsupervised image captioning with shared multimodal embeddings," *CoRR*, vol. abs/1908.09317, 2019.

[18] Wenliang Dai, Samuel Cahyawijaya, Zihan Liu, and Pascale Fung, "Multimodal end-to-end sparse model for emotion recognition," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online, June 2021, pp. 5305–5316, Association for Computational Linguistics.

[19] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, "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: Long Papers)*, Melbourne, Australia, July 2018, pp. 2236–2246, Association for Computational Linguistics.

[20] Chuanhe Liu, Wenqiang Jiang, Minghao Wang, and Tianhao Tang, "Group level audio-video emotion recognition using hybrid networks," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 807–812, Association for Computing Machinery.

[21] Yanan Wang, Jianming Wu, Panikos Heracleous, Shinya Wada, Rui Kimura, and Satoshi Kurihara, "Implicit knowledge injectable cross attention audiovisual model for group emotion recognition," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 827–834, Association for Computing Machinery.

[22] Sandra Ottl, Shahin Amiriparian, Maurice Gerczuk, Vincent Karas, and Björn Schuller, "Group-level speech emotion recognition utilising deep spectrum features," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 821–826, Association for Computing Machinery.

[23] Shivam Srivastava, Saandeep Aathreya SIdhapur Lakshminarayan, Saurabh Hinduja, Sk Rahatul Jannat, Hamza Elhamdadi, and Shaun Canavan, "Recognizing emotion in the wild using multimodal data," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 849–857, Association for Computing Machinery.

[24] Mo Sun, Jian Li, Hui Feng, Wei Gou, Haifeng Shen, Jian Tang, Yi Yang, and Jieping Ye, "Multi-modal fusion using spatio-temporal and static features for group emotion recognition," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, New York, NY, USA, 2020, ICMI '20, p. 835–840, Association for Computing Machinery.

[25] João Ribeiro Pinto, Tiago Gonçalves, Carolina Pinto, Luís Sanhudo, Joaquim Fonseca, Filipe Gonçalves, Pedro Carvalho, and Jaime S. Cardoso, "Audiovisual classification of group emotion valence using activity recognition networks," in *2020 IEEE 4th International Conference on Image Processing, Applications and Systems (IPAS)*, 2020, pp. 114–119.

[26] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li, "Bag of tricks for image classification with convolutional neural networks," 2018.

[27] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," *CoRR*, vol. abs/1608.00859, 2016.

[28] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatiotemporal representation with pseudo-3d residual networks," *CoRR*, vol. abs/1711.10305, 2017.

[29] Dan Hendrycks and Kevin Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *CoRR*, vol. abs/1606.08415, 2016.

[30] Dongyoon Han, Sangdoo Yun, Byeongho Heo, and YoungJoon Yoo, "Rexnet: Diminishing representational bottleneck on convolutional neural network," 2020.

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

[32] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021.

[33] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[35] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018.

[36] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

[37] Weiyao Wang, Du Tran, and Matt Feiszli, "What makes training multi-modal networks hard?," *CoRR*, vol. abs/1905.12681, 2019.