Audio-Based Emotion Recognition Enhancement Through Progressive GANS

AUDIO-BASED EMOTION RECOGNITION ENHANCEMENT THROUGH PROGRESSIVE GANS

Christos Athanasiadis, Enrique Hortal and Stylianos Asteriadis

Maastricht University, Department of Data Science and Knowledge Engineering (DKE), the Netherlands

ABSTRACT

Training large-scale architectures such as Generative Adversarial Networks (GANs) in order to investigate audio-visual relations in emotion-enriched interactions is a challenging task. This procedure is hindered by the high complexity as well as the mode collapse phenomenon. Sufficiently training these architectures requires a massive amount of data. Furthermore, creating extensive audio-visual datasets for specific tasks, like emotion recognition, is a complicated task handicapped by the annotation cost and labelling ambiguities. On the other hand, it is much more forthright to get access to unlabeled audio-visual datasets mainly due to the easy access to online multimedia content. In this work, a progressive process for training GANs was conducted. The first step leverages enormous audio-visual unlabeled datasets to expose concealed cross-modal relationships. Meanwhile in the second step, a calibration of the weights by employing a limited amount of emotion annotated data was performed. Through experimentation, it was shown that our progressive GANs schema leads to a more efficient optimization of the whole network, and the generated samples from the target domain, when fused with the authentic ones, provides enhanced emotion recognition results.

Index Terms— Domain Adaptation, Affective Computing, Generative Adversarial Networks

1. INTRODUCTION

Emotion recognition through facial expressions is a field that has been extensively studied from the affect computing perspective [1] [2]. However, the progress regarding emotion recognition through audio modality is not yet as advanced [3]. A possible explanation for this is the reduced number of publicly available databases for audio emotion recognition in contrast with the availability of datasets in the case of face modality [3]. That being said, the engineering of such big and complex databases is not always a forthright task. In order to address these limitations, domain adaptation (DA) is fostered in the current work in an effort to learn a projection between face and audio domains. At the theoretical level, the current work was inspired by research conducted in the domain of cognitive psychology [4][5] where audio-visual relationships were studied, and more particularly, within the emotional context [6]. We intended to conduct analogous research, but from the affective computing perspective, and to investigate whether it is possible to transfer knowledge between facial expressions to related audio data.

In this work, GANs [7] were introduced and examined as a candidate solution for learning the projection between source and target domains. The objective was to generate samples from the target domain (audio) given as an input the source domain (face). That was done in order to expand existing datasets, with generated samples, in an effort to extend annotated audio data and perform more sophisticated emotion recognition.

In particular, inspired by the work done in [8] and [9], a deviation from the initial algorithm that performed image-to-image translation was introduced. In order to take advantage of this type of translation, we decided to use, as audio features, spectrograms (which stands for the frequency fluctuations over time). Concerning the face modality, face detection and alignment was performed for the frames of each video and finally the middle frame for each video was used. The modified GANs architecture that we applied is portrayed in Figure 1. In this altered version, given as input the face modality, some conditional emotion-related information and a random noise vector, the approach generates samples that are distributed in the audio domain (but maintaining the emotion-related information). In this approach, there is an extra network added to the whole architecture in comparison with the initial version of GANs, the so-called "Classifier" (Q). This network aims to better calibrate the generator to produce samples that come from specific emotional states.

However, directly learning this aforesaid projection is still a very complicated procedure that requires sophisticated deep architectures and access to enormous datasets that are also consisted of linked annotations related to the emotion states of the subjects from videos. For this reason, a progressive training of the whole architecture is proposed in order to improve the learning capacity of the whole framework. This method consists of two steps. Insofar as the first step is concerned, the so-called "weight initialization", a huge data corpus of unlabeled audio-visual clips, derived from the VoxCeleb dataset [10], is used in an attempt to uncover low-level hidden relationships between the face and audio modalities.

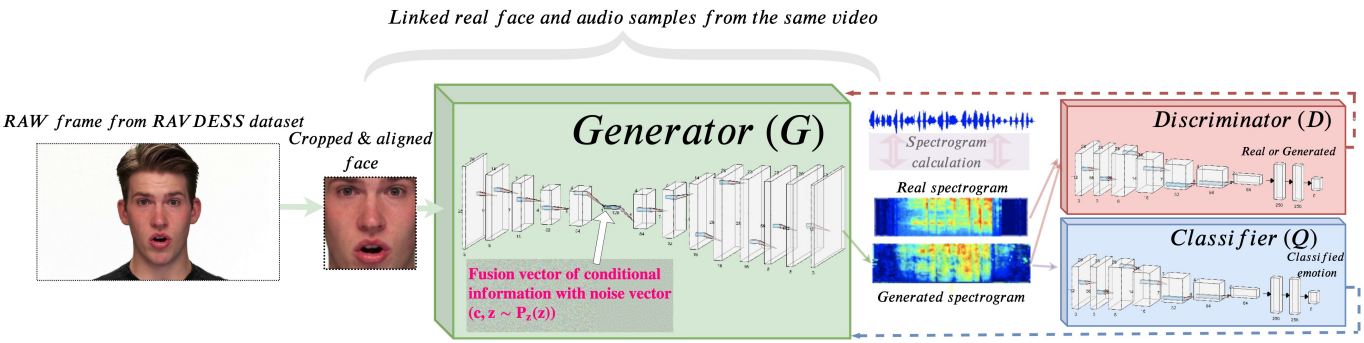


Fig. 1: The complete architecture of the whole approach and the pre-process step for the face and audio samples.

As a step further, for the sake of calibrating the network in an emotion-wise manner, we made use of emotion-labeled datasets, namely CREMA-D [11] and RAVDESS [12].

The current framework and the adopted architecture was prompted from the work done in [13] where a novel CNN was introduced. Another fertile inspiration was, the work done in [14], where authors attempted to transfer knowledge (between different visual datasets) from pre-trained GANs networks to new domains, as a fine tuning step. By contrast, our approach was performed for cross-modal GANs which were assembled from different networks. Finally, the conducted research of this paper is the extension of the work done in [15]. The difference from the previous paper is two-fold: 1) the implementation of the knowledge transfer process occurred by using a progressive calibration of the weights and 2) a more wealth evaluation scheme was applied to measure the quality of the approach (by measuring the Structural Similarity Index [19]).

2. APPROACH

The overall architecture of the approach can be seen in Figure 1, which shows our methodology for investigating the audio-visual cross-modal relationships. In the initial version of GANs, generator G and discriminator D networks work in tandem playing a min-max game in an effort to learn the target distribution domain. In particular, this can be outlined using the following equation:

$$\min_G \max_D F_1(D,G) = E_{y\sim X_T(x)}[\log D(x)] + E_{z\sim P_z}[\log(1 - D(G(z)))] \quad (1)$$

where $y \in X_T$ are the data from the distribution that we want to approximate and $z \in P_z$ corresponds to the uniform noise vector. Nonetheless, our intention in the current work is to employ a cross-domain projection between the face and audio domains. On this ground, the input to the network G is altered to be the source domain $X_S = \{x_1, x_2, ..., x_n\} \subseteq D_S$ and the noise vector $z \in P_z$ likewise [15][16] (Figure 1 visualize the input to G). Thereupon, Equation 1 can be transformed into:

$$\min_G \max_D F_2(D,G) = E_{y\sim X_T}[\log D(y)] + E_{z\sim P_z, x\sim X_S}[\log(1 - D(G(x,z)))] \quad (2)$$

Additionally, since the goal is to generate data that approximate the target domain $X_T \subseteq D_T$ conditioned to emotional information as in [22], Equation 2 could be easily re-shaped as:

$$\min_G \max_D F_3(D,G) = E_{y\sim X_T}[\log D(y)] + E_{z\sim P_z, x\sim X_S}[\log(1 - D(G(x|c,z)))] \quad (3)$$

where the input of the G network is conditioned to the variable information c. In the current framework, we examined two alternative sources of conditional emotion-related information: the label information provided in the datasets and the prediction of a classifier trained using datasets derived from the source domain X_S. On top of that, similarly with [17], we explored the possibility of fusing the initial GANs objective with a pixel-wise loss, the $L_1(G(x|c,z))$ distance [18] (for our case). By fusing that loss in Equation 3 we can derive to:

$$\min_G \max_D (V_3(D, G(x|c,z)) + L_1(G(x|c,z))) \quad (4)$$

So far, we made use of the conditional variable c. To increase efficiency, an extra network $Q = f_T(x_{gen.} \in G(x|c,z))$ that calculates an error based on the correct or wrong classification of the emotional states (x_{gen} represents the generated samples) was added.

The proposed network Q is a CNN network with an architecture similar to the network D (see Figure 1). The input of this network is the output samples of G and the calculated error of this network was passed to the Generator optimization in tandem with Formula 4. In our approach, the intro-

duced error of the classifier, the cross-entropy of the generated samples from the network G, could be denoted as: $Q = E[\log(P(x_k|D(G(x|c,z)))]$, where x_k indicates the probability of a sample to derive from class k. Finally, the complete loss function is the summary of Formula 4 and the aforementioned cross-entropy loss.

However, directly training the proposed architecture is a notably difficult task. Furthermore, it was observed, through experimentation, that the proposed architecture needs to be complex enough to learn a meaningful projection between both domains. In total, during the training procedure, our architecture needs to calibrate more than 50 million parameters. Directly learning this number of parameters requires an extremely huge data corpus of correlated face and audio samples annotated with the emotional states. Unfortunately, such a big dataset, with these specifications, could not be easily met in the affective computing. To facilitate this task, a **progressive learning** procedure was introduced.

This approach was proposed in an effort to better tune the weights of the architecture and increase the efficiency of the approach. It consists of two steps: Firstly, an enormous corpus of unlabeled data [10] is utilized with the purpose of initializing the weights of the networks G and D. This is done in an attempt to map and uncover the relationship between the two domains (X_S and X_T) without placing any additional conditional constraints. The defined task is to fine-tune a network G to generate samples from the target domain conditioned merely to the low-level pixel information of the source domain X_S. In this manner, low-level, not related to emotions, correlations among features can be retrieved and pave the path for domain adaptation at a later stage. Then, having initialized the weights of G to produce unlabelled data from the target domain, the next step is to continue the optimization of the whole network by using a fully annotated audio-visual dataset with a view to calibrating the weights and to expose high-level emotion-wise relationships between face and audio domain.

3. EXPERIMENTAL PHASE

3.1. Experimental protocol

The core objective of this work is two-fold: 1) to prove the efficiency of the **progressive training** approach; and 2) to evaluate the capacity and the amount of knowledge transferred between the source and target domains. For both objectives, the accuracy of emotion classification by using only real audio samples was established as our baseline evaluation (from CREMA-D and RAVDESS). This evaluation was compared with the performance when blending these data with generated audio samples. The classifier employed to establish this baseline is a network similar to the one used in our GANs approach (represented in Figure 1 as "Classifier Q").

Furthermore, except of the classification evaluation schema, structural Similarity Index (SSIM) [19] was adopted as the main evaluation tool with respect to the quality of the generated results. In such a manner, we can measure the visual fidelity of the samples and the wealthiness in emotion reactions. That approach, allows to compare two images based on perceptual differences. It is expressed as a floating-point number ranging from -1 to 1, where 1 indicates exact similarity and -1 means a complete dissimilarity. The comparison is based on the following three measurements (between the two data samples): luminance, contrast and structure. Furthermore, we employed FID [20] and IS [21] metrics to validate the results of SSIM.

Finally, for evaluating the second objective, progressive optimization of the whole process, we implemented our whole GANs scheme with (**progressive GANs**) and without (**moderate GANs**) the utilization of the initialization process.

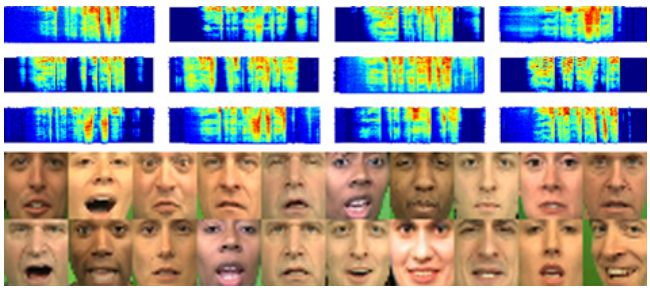
For the **moderate training** the whole approach was evaluated using two scenarios: supervised and semi-supervised. Firstly, the whole approach was trained in a **supervised** manner, by using as the conditional variable c the label information provided from the datasets. Secondly, the label information that was given as input to the generation G was replaced with the output of a classifier in the source domain (faces) $f_S(X_S)$ (that has the same architecture like the Q network) to perform emotion recognition on the target domain (**semi-supervised**). The output of this classifier was a six-featured vector that contains, in each feature, the probability of the input sample to be derived from a specific emotional label.

Finally, in the case of **progressive training**, the same scenarios were tested. However, the core difference is the preliminary progressive optimization of the weights performed by utilizing the unlabeled VoxCeleb dataset. That was done in an effort to show whether it is possible to increase our approach efficiency with the progressive optimization procedure. Additionally, merely for visualization purposes, further experiments were performed by switching the domains and therefore, generating new faces (target domain) considering audio as the source domain. In this case, the architecture presented in Figure 1 was slightly modified to fit the needs of this projection. Furthermore, in contrast with the work done in [15] a more sophisticated network G was introduced in this work to facilitate the needs of the **progressive training** and capture the nuance patterns of the VoxCeleb dataset.

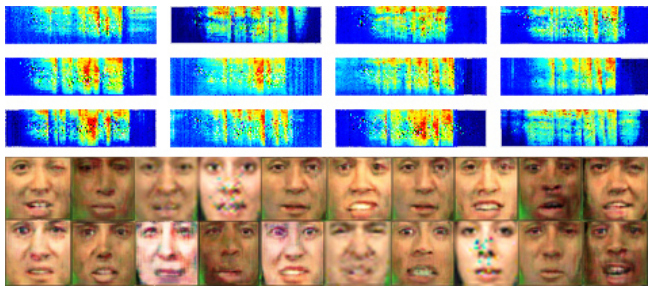
3.2. Experimental results and discussion

In Table 1, the results for all the aforementioned evaluated cases are encompassed. Firstly, the results for our reference metric could be seen, denoted in the Table as "Baseline".

In the light of the results obtained, it can be observed that, when the generated samples get fused together with the initial ones, we managed to obtain an improvement in the emotion recognition performances. Another notable remark was that, while the results in all **semi-supervised** scenarios were in-



(a) Real faces (CREMA-D) and spectrograms (RAVDESS).



(b) Generated faces and spectrograms.

Fig. 2: Samples derived from CREMA-D and RAVDESS.

Table 1: Classification performance, FID and IS for all the methods analysed.

Case	CREMA-D				RAVDESS			
	Classification	FID	IS	SSIM	Classification	FID	IS	SSIM
Baseline	49.34%				44.73%			
Mod.Supervised GANs	52.52%	59.88	2.25	0.82	47.11%	49.77	2.01	0.93
Mod.semi-Supervise GANs	49.92%	60.13	2.21	0.80	46.23%	48.65	1.98	0.94
Pr.Supervised GANs	**53.71%**	**59.60**	**2.31**	**0.85**	**47.55%**	**47.95**	**2.13**	**0.95**
Pr.semi-Supervised GANs	50.40%	59.91	2.22	0.81	46.77%	48.98	2.05	0.95

ferior than in the case of **supervised** (see Table 1), they were still marginally better than the baseline for both datasets. This shows that we can train our framework even if there is limited access to annotated data for the target task by including an additional classifier in a semi-supervised fashion.

From the visual results obtained for CREMA-D and RAVDESS (Figure 2b), it was deduced that the proposed approach, in all steps, managed to generate visual results that are considered faithful representations of the face and audio domains and also incorporate the emotional context (that is much easier to inspect in the generated faces). As it is rendered in Table 1, the best obtained results that connote the efficiency of our progressive approach were established for the CREMA-D and RAVDESS datasets, in the supervised scenario validated by FID and IS as well. These metrics highlighted that the quality and the distributions of generated images during the **progressive training** of GANs could be closer to the initial image distribution rather than the case of the **moderate training**. For the SSIM, while for the CREMA-D dataset we obtained the expected behaviour, for RAVDESS the improvement between moderate and progressive training cases was marginal.

A final evaluation was performed by removing from the whole architecture the network Q and then proceeding with the aforementioned scenarios. In all these cases, we noticed that our methodology was not able to sufficiently generate spectrograms and the fused datasets led to marginally worse results than the baseline ("negative knowledge transfer") both for classification and visual quality metrics.

4. CONCLUSION

In the current work, we investigated the research question of whether the progressive training of our GANs framework could improve the audio-visual cross-modal transfer knowledge. Our aim was to uncover an audio-visual projection by progressively optimizing the weights of the GANs architecture. That was done by firstly exposing low-level relationships between face and audio which are not related to emotional states by employing an unlabelled corpus of audio-visual data. Subsequently, the architecture's weights were trained using smaller emotion-related datasets with a view to calibrating the weights and to expose high-level emotion-wise relationships between the face and audio domains. The conducted evaluation scheme was affirmative to the progressive training assumption where an increased cross-modal knowledge transfer (in comparison with the moderate training) during emotion-enriched interactions was obtained. This comparison was established by a two-fold validation: firstly based on the classification performance (as it is rendered in Table 1). Secondly, it was shown that progressive training can significantly improve the quality of the generated data (measured by SSIM, IS and FID), especially when the volume of annotated data is limited.

5. ACKNOWLEDGEMENTS

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Nvidia GeForce GTX Titan XP GPU used throughout the experimental phase.

6. REFERENCES

[1] C.A.Corneanu, M.Oliu, J.F.Cohn and S.Escalera, A Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition:History, Trends, and Affect related Applications, IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 38, Number 8, Pages: 1548–1568, 2016

[2] J.Kumari, R.Rajesh and K.M.Pooja, Facial Expression Recognition: A Survey, Procedia Computer Science Volume 58, Pages 486-491, 2015.

[3] S.Albanie, A.Nagrani, A.Vedaldi and A.Zisserman, Emotion Recognition in Speech using Cross-Modal Transfer in the Wild, ACM Multimedia 2018.

[4] T.Grossman, The development of emotion perception in face and voice during infancy, Resorative Neurology and Neuroscience 28, Pages: 219–236, 2010.

[5] D.W.Massaro and M.M.Cohen, Perceiving Talking Faces, Current Directions in Psychological Science, Volume 4, Number 4, 1995.

[6] M.D.Pell, Prosody–face Interactions in Emotional Processing as Revealed by the Facial Affect Decision Task, Journal of Nonverbal Behavior, Volume 29, Number 4, Pages: 193-–215, 2005.

[7] I.J.Goodfellow, J.P.-Abadie, M.Mirza, B.Xu, D.W.Farley, S. Ozair, A.Courville and Y. Bengio, Generative Adversarial Networks, 27th conference on Advances in Neural Information Processing Systems (NIPS), 2014.

[8] P.Isola, J.Yan, Z.Tinghui, Z.Alexei and A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[9] T.Kim, M.Cha, H.Kim, J.K.Lee and J.Kim, Learning to Discover Cross-Domain Relations with Generative Adversarial Networks, International Conference on Machine Learning (ICML), 2017.

[10] A.Nagrani, J.S.Chung and A.Zisserman, VoxCeleb: a large-scale speaker identification dataset, in Interspeech, 2017.

[11] H.Cao, D.G.Cooper, M.K.Keutmann, R.C.Gur, A.Nenkova and R.Verma, CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset, IEEE Transactions on Affective Computing, Volume: 5, Number: 4, Pages: 377–390, 2014.

[12] S.R.Livingstone and F.A.Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, Volume 13, Number 5, Pages: 1—35, 2018.

[13] O.Ronneberger, P.Fischer and T.Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation, International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[14] Y.Wang, C.Wu, L.Herranz, J.V.D.Weijer, A.G.Garcia and B.Raducanu, Transferring GANs: generating images from limited data, European Conference on Computer Vision (ECCV), 2018.

[15] C.Athanasiadis, E.Hortal and Stylianos Asteriadis, Audio–visual domain adaptation using conditional semi-supervised Generative Adversarial Networks, Elsevier Neurocomputing, 2019.

[16] X.Wang and A.Gupta. Generative image modeling using style and structure adversarial networks, European COnfeence on Computer Vision (ECCV), 2016.

[17] P.Isola, J.Yan, Z.Tinghui, Z.Alexei and A. Efros, Image-to-Image Translation with Conditional Adversarial Networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[18] D.Pathak, P.Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[19] Z.Wang, A.C.Bovik, H.R.Sheikh and E.P.Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, EEE Transactions on image processing, Volume 13, Number 4, 2004.

[20] M.Heusel, H.Ramsauer, T.Unterthiner, B.Nessler and S.Hochreiter, Gans trained by a two time-scale update rule converge to a local Nash equilibrium, In Advances in Neural Information Processing Systems (NIPS), (2017).

[21] T.Salimans, I.Goodfellow, W.Zaremba, V.Cheung, A.Radford and X.Chen, Improved Techniques for Training GANs, In Advances in Neural Information Processing Systems (NIPS), 2016.

[22] M.Mirza and S.Osindero, Conditional Generative Adversarial Nets, Computing Research Repository (CoRR), 2014.