In SSR 2018 proceedings, International Ph.D. Conference on Safe and Social Robotics (SSR), pp. 49--52.
Madrid (Spain), 29-30 September 2018, http://www.socrates-project.eu/sesoro-2018/.

An Adaptive Neural Approach Based on Ensemble and Multitask Learning for Affect Recognition

Henrique Siqueira

***Abstract*—In this paper, we evaluate the effect of Multitask Learning (MTL) in an ensemble with shared representations based on convolutional networks in the task of affect recognition from facial expressions. Our convolutional architecture is divided into three levels of hierarchy regarding MTL. The first level is conditioned to learn lower-level representations, which are shared with independent convolutional branches related to different tasks on the second level. While each independent branch is fostered to learn task-specific representations, the early shared layers are fostered to learn features that are relevant to multiple tasks due to the inductive transfer mechanism from MTL. The third level consists of an ensemble of convolutional branches responsible for learning higher-level representations and allowing re-training with unlabelled expressions. Our experiments show a slight improvement in recognition performance using MTL over Single Task Learning (STL) on the AffectNet dataset, but a significant reduction in training time. Finally, we discuss the potential use of MTL and hard constraints into the inference and re-training processes of the proposed approach to improve its generalization performance.**

***Index Terms*—Semi-supervised Learning, Multitask Learning, Ensemble Methods, Facial Expression Recognition**

I. Introduction

With the advance in health care, the modern society is enjoying longer lives. The long life expectancy accompanied by low birth rates dictate the growth of ageing populations in several countries, which already comprise over a tenth of the global population [1]. Besides physical health, psychological and sociological factors have a significant impact on well-being and good life quality in old age. Sociability, in particular, plays a crucial role against loneliness in advanced years, which is one of the main factors that lead older adults to experience feelings of depression and thoughts of mortality [2].

Studies from different areas including robotics, medicine and economics have suggested making use of social robots as home companions and social assistants in senior care facilities to address loneliness among older adults and to support their needs and independence [1]. In addition to their functional activities (e.g., dispensing of medication and providing reminders), such robots can establish social and affective relationships with older adults which reduce feelings of loneliness among older people and provide warm caregiving to them, as investigated by Pols and Moser [3].

A fundamental aspect of social robots is their affective capabilities; the ability to recognize, express or even have emotions, albeit having simulated ones [4]. Emotions are highly present in human interactions, by influencing our rational thinking and decision-making [5]. Sad facial expressions and a low tone of voice during a conversation, for instance, might encourage a friend to comfort you [6]. A social robot capable of identifying and using this emotional information for making decisions could enhance its social skills by initiating an interaction with a senior perceived as sad to support them with positive messages. As evidenced by Sabelli et al. [7] through an ethnographic study of a conversational agent in an elderly care center, such emotionally support not only improve engagement in interacting with a robot, but also reduce loneliness and positively regulates their feelings.

Despite the remarkable progress in the area of automatic emotion recognition (see Poria et al. [9] for a recent review of affective computing), most of the existing approaches are extensively trained using supervised learning techniques on a given dataset [10], [11], which frequently drop in recognition performance when trialled under different conditions than the one used for training [12], [13]. Taylor et al. [14] suggested that this drop in recognition performance may be caused by the inability of those approaches to account for individual differences, since the same emotional state can be expressed differently among individuals [5], [15]. Even the same person may present a high physiological variation for the same emotional state in different days [16]. Therefore, an emotion recognition system that could improve recognition performance over time with unlabelled expressions is beneficial to social robots as they could be able to enhance their emotional capabilities over interactions. This adaptive capability is especially needed for social robots in senior care facilities, since emotional expression variations in older adults may be even higher due to cognitive or physical issues [17].

As investigated in our previous work [8], an ensemble with shared representations can potentially be used as an adaptive emotion recognition system for social robots, where emotional expressions collected from human-robot interactions can be utilized for re-training the ensemble. Although re-training the system using the ensemble predictions from unlabelled expressions led to an improvement in recognition performance in the majority of cases, there were few cases where it degenerated the recognition capability. We hypothesize that providing more information about an emotional expression via Multitask Learning (MTL) might not only yield to better generalization performance, but might also make the re-training phase through ensemble predictions more efficient. MTL can be defined as an inductive transfer learning mechanism where multiple *related* tasks are trained in parallel using shared

The author is with Knowledge Technology, Department of Informatics, University of Hamburg, Vogt-Koelln-Str. 30, 22527 Hamburg, Germany siqueira@informatik.uni-hamburg.de

In SSR 2018 proceedings, International Ph.D. Conference on Safe and Social Robotics (SSR), pp. 49--52.
Madrid (Spain), 29-30 September 2018, http://www.socrates-project.eu/sesoro-2018/.

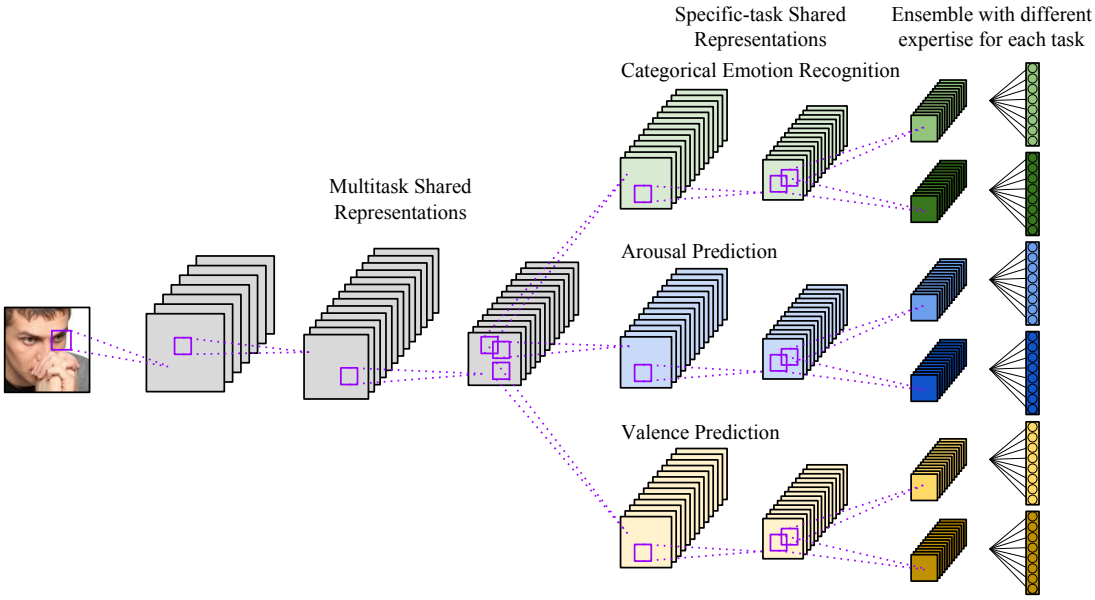

Fig. 1. Illustration of the proposed architecture for multitask learning. While the early layers in gray learn lower-level representations useful for multiple tasks, the three separate convolutional branches in green (top), blue (middle) and yellow (bottom) are employed to learn task-specific representations. On the right, the ensemble of convolutional branches is adopted as proposed by Siqueira et al. [8] for learning higher-level representations for each task.

representations [18]. Several studies have demonstrated the benefits of multitask learning on improving generalization performance and decreasing training time in contrast to Single Task Learning (STL) [14], [18], [19], where a machine learning method learns only one task at a time. Devries et al. [19] have demonstrated that facial expression recognition can be improved by training a convolutional neural network to detect facial landmarks as an auxiliary task in an MTL setting. Taylor et al. [14] employed MTL to account for individual differences for mood prediction by first clustering them regarding personality and gender, and subsequently, each cluster considered as a different prediction task, which resulted in an overall improvement on the generalization performance.

In this paper, we adapt our previous approach to employ multitask learning. Our approach consists of designing a convolutional architecture based upon three different levels of hierarchy regarding MTL. The first level is responsible for learning lower-level representations from the data. These representations are shared between multiple and independent branches in the second level, where each branch is constrained to learn features relevant to a particular task. In this work, we consider as related tasks the recognition of categorical emotional expressions (e.g., happy, sad and neutral), and the prediction of arousal and valence levels from the dimensional representations of emotion by Russell [20]. Lastly, the third level is an ensemble of convolutional branches with different expertise for each task, as described in Siqueira et al. [8]. In addition of presenting our preliminary analysis of the effect of MTL on the generalization performance on the AffectNet dataset [12], we discuss the potential benefits of using multiple information from the same input as hard constraint [21] in the inference and re-training processes of the proposed approach.

II. Approach

In the proposed approach illustrated in Figure 1, the early convolutional layers learn lower-level representations from the training data. They are conditioned to discovery features that are suitable to different and related tasks by the inductive transfer learning from multiple teach signals which are back-propagated from each task-related output to the shared representations, as defined by Caruana [18]: *"the multitask bias causes the inductive learner to prefer hypotheses that explain more than one task"*. These lower-level representations are shared between independent convolutional branches, each related to a specific task represented by different colors in Figure 1. The green convolutional branch, in the context of this paper, is fostered to learn important features to distinguish categorical emotions, where the blue and yellow branches to learn relevant features for predicting arousal and valence, respectively. In the highest level of the architecture, an ensemble of convolutional branches is employed as proposed in the work of Siqueira et al. [8]. The major goal for each branch in the ensemble is the development of higher-level representations from the training data that are different and complementary to other branches' expertise. If this assumption is satisfied, recognition performance might be improved by re-training the ensemble with their own predictions [8].

While multitask learning may improve the generalization capability of a model by fostering shared layers to learn features that are useful for different tasks, the different pieces of information gathered for each task from the same emotional expression might provide supplementary evidence for the correct classification of such expression. As an example, suppose that the convolutional branches responsible for the categorical emotion recognition classify a given expression, with a certain

In SSR 2018 proceedings, International Ph.D. Conference on Safe and Social Robotics (SSR), pp. 49--52.
Madrid (Spain), 29-30 September 2018, http://www.socrates-project.eu/sesoro-2018/.

degree of uncertainty, as happy or sad. Uncertainty cases may occur when some branches classify an image as belonging to a class A, while other branches classify the same image as belonging to a class B. By using prior knowledge about the task and different pieces of information from the same input, the valence prediction could have charged the same expression in our example as positive, and hence, the confidence for the categorical emotion recognition could have been increased towards the happy category. This strategy can be understood as imposing hard constraints in the inference and training processes, and this field of study is well explored in the book of Gori [21]. In spite of the potential benefits of imposing hard constraints into our approach, the experiments conducted for this paper are limited to the analysis of the effect of MTL on the recognition performance.

III. Preliminary Experiments

Fig. 2. Examples of the eight discrete categories from the AffectNet dataset [12] adopted in our experiments: Neutral (Ne), Happy (Ha), Sad (Sa), Surprise (Su), Disgust (Di), Fear (Fe), Anger (An) and Contempt (Co).

We evaluated the proposed approach on the AffectNet dataset [12], which consists of over a million face images collected by querying search engines with emotion-related keywords in six different languages. AffectNet is divided into the labelled training, unlabelled training, validation and test sets. Each set was manually annotated in terms of categorical and dimensional representations of emotion, except the unlabelled training set. In addition to the universal facial expressions proposed by Ekman [22] (see Figure 2), such as Happy (Ha), Sad (Sa), Surprise (Su), Fear (Fe), Disgust (Di), Anger (An) and Contempt (Co), the categorical representation of AffectNet also presents Neutral (Ne), None (No), Uncertain (Un) and Non-Face (NF) categories. For the dimensional representation, the dataset was annotated based on the circumplex model of affect proposed by Russell [20], where the *arousal* level indicates how excited or calm an event is, the *valence* level indicates how pleasant or unpleasant an event is. Continuous values ranging from -1 to 1 were assigned to emotional facial expressions, whereas -2 indicates images that belong to non-face and uncertain categories.

Our architecture is divided into three levels. The first level consists of three convolutional layers with 64, 128 and 256 filters. These lower-level representations are shared between three convolutional branches, one for each task: the classification of categorical emotions, and the prediction of arousal and valence levels. Each convolutional branch has one convolutional layer with 512 filters for learning features relevant to a specific task. Until this level, all of the convolutional layers are followed by batch normalization and max-pooling layers with a pool size of 2. The third and highest level is an ensemble of convolutional branches. For the categorical emotion recognition task, four branches compose the ensemble. Each branch in the ensemble is composed of one convolutional layer with 1024 filters, followed by the global average pooling layer, and the output layer with 8 neurons. To foster the development of different and complementary features for the same task in the ensemble, a different weighted loss function is assigned for each branch. This overall configuration is also adopted in the other two branches, which are responsible for predicting arousal and valence levels. However, their output layers have 41 neurons each, representing the discrete counterpart of the continuous emotional scales. This discretisation is necessary to assign a unique weighted loss function for each branch. As activation function, ReLU is adopted for all of the neurons, except the output layer where the softmax function is applied. During validation, we take the mean probability distribution from the ensemble.

TABLE I
Accuracy (%) and RMSE on AffectNet for categorical and dimensional representations of emotion.

Approaches	Categorical	Arousal	Valence	Params
MTL	50.32%	**0.37**	0.46	**50M**
STL	48.05%	0.39	0.47	**50M**
Mollahosseini et al. [12]	**58.00%**	0.41	**0.37**	180M

We adopt the single task learning counterparts of the proposed architecture as baselines. Thus, the network trained for categorical emotion recognition consists of five convolution layers with 64, 128, 256 and 512 filters, followed by an ensemble with four convolutional branches, each of which consisting of a convolutional layer with 1024 filters, an average pooling layer, and an output layer with 8 neurons. In addition to the comparisons with the baseline networks, we also compare our results with the approach proposed by Mollahosseini et al. [12] in the AffectNet paper. In their work, three different AlexNets [23] were re-trained on the AffectNet dataset, outperforming traditional classifiers and off-the-shelf facial expression recognition systems such as support vector machines and Microsoft Cognitive Services emotion API [1]. The faces are cropped using the facial coordinates provided by the dataset, and re-scaled to 96 x 96 pixels to reduce the computational cost. The pixel intensities from each image are normalized between 0 and 1. The networks were trained for 15 epochs using RMSProp with an initial learning rate of 0.001.

A. Initial Results and Discussion

Table I shows the accuracy for the categorical classification of emotions, the root-mean-square error (RMSE) for the predictions of arousal and valence levels, and the number of trainable parameters for each approach. MTL represents the proposed approach trained for multiple related tasks in parallel, whereas STL represents its counterpart but trained for one task at a time. Therefore, the results reported for STL are three different convolutional networks trained from scratch

[1]https://www.microsoft.com/cognitiveservices/enus/emotionapi

In SSR 2018 proceedings, International Ph.D. Conference on Safe and Social Robotics (SSR), pp. 49--52.
Madrid (Spain), 29-30 September 2018, http://www.socrates-project.eu/sesoro-2018/.

on AffectNet. This is also true for the results reported by Mollahosseini et al. [12]. Each AlexNet re-trained by them has roughly 60 million trainable parameters [23], resulting in 180 million parameters for the three networks.

Although the recognition performance of MTL and STL are similar, with the first reaching slight higher accuracy for categorical emotion classification, and lower RMSE for the arousal and valence predictions, the proposed approach can be trained t times faster than STL, being t the number of tasks to be learnt. The training time factor might be crucial for the application of the proposed approach for continual learning in robotic platforms, especially robots with limited computational resources. When compared with the methods proposed by Mollahosseini et al. [12], the proposed approach has achieved a substantial lower RMSE for arousal prediction, but has presented an inferior performance for categorical emotion classification and valence prediction. However, the adaptation of their methods for continual learning, where a robot should improve its recognition performance over time might be infeasible due to the high number of parameters.

IV. Conclusions and Future Work

We adapted our previous work on an ensemble with shared representation to account for multitask learning. MTL acts as an inductive transfer learning mechanism that frequently improves generalization performance by fostering shared representations to learn features that are useful for different tasks. Although the employment of multitask learning provided a small gain in recognition performance, it provided a significant reduction in training time since several tasks can be trained in parallel. This training time gain is an important factor for continual learning in social robots, since response time is fundamental to a natural interaction. Moreover, we discussed how different pieces of information from the same input regarding MTL could be used as hard constraints in the inference and training processes for improving generalization performance.

As future work, we will analyse the internal representations related to each level of hierarchy regarding MTL in the proposed approach. This analysis might explain the slight improvement on generalization performance obtained in our experiments. Furthermore, the potentiality of MLT and hard constraints for improving generalization performance discussed in this paper will be evaluated on AffectNet, including an analysis of the adaptive behaviour of the proposed approach on the unlabelled training set. In addition to a static dataset of emotions, the proposed approach will also be evaluated in a more naturalistic condition, where not only spatial but also temporal features are presented in the expression of the individual emotional state [24].

V. Acknowledgement

This work has received funding from the European Unions Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721619 (SOCRATES).

References

[1] I. Pedersen, S. Reid, and K. Aspevig, "Developing social robots for aging populations: A literature review of recent academic sources," *Sociology Compass*, vol. 12, no. 6, p. e12585, 2018.

[2] A. Singh and N. Misra, "Loneliness, depression and sociability in old age," *Industrial psychiatry journal*, vol. 18, no. 1, p. 51, 2009.

[3] J. Pols and I. Moser, "Cold technologies versus warm care? on affective and social relations with and through care technologies," *ALTER*, vol. 3, no. 2, pp. 159–178, 2009.

[4] L.-F. Rodríguez and F. Ramos, "Computational models of emotions for autonomous agents: major challenges," *Artificial Intelligence Review*, vol. 43, no. 3, pp. 437–465, 2015.

[5] R. W. Picard, "Affective computing," MIT Media Laboratory, Perceptual Computing, Tech. Rep., 1997.

[6] R. Kirby, J. Forlizzi, and R. Simmons, "Affective social robots," *Robotics and Autonomous Systems*, vol. 58, no. 3, pp. 322–332, 2010.

[7] A. M. Sabelli, T. Kanda, and N. Hagita, "A conversational robot in an elderly care center: An ethnographic study," in *2011 6th ACM/IEEE International Conference on HRI*, March 2011, pp. 37–44.

[8] H. Siqueira, P. Barros, S. Magg, and S. Wermter, "An ensemble with shared representations based on convolutional networks for continually learning facial expressions," in *Accepted in Intelligent Robots and Systems (IROS) IEEE/RSJ International Conference. IEEE.*, 2018.

[9] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[10] P. Khorrami, T. L. Paine, and T. S. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" in *Proceedings of the 2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, ser. ICCVW '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 19–27.

[11] P. Barros, C. Weber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, Nov 2015, pp. 582–587.

[12] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, pp. 1–1, 2017.

[13] H. Siqueira, P. Barros, S. Magg, C. Weber, and S. Wermter, "A sublayered hierarchical pyramidal neural architecture for facial expression recognition," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, Apr 2018, pp. 1–6.

[14] S. A. Taylor, N. Jaques, E. Nosakhare, A. Sano, and R. Picard, "Personalized multitask learning for predicting tomorrow's mood, stress, and health," *IEEE Transactions on Affective Computing*, no. 1, pp. 1–1, 2017.

[15] A. Siqueira, Henrique Sutherland and, P. Barros, M. Kerzel, S. Magg, and S. Wermter, "Disambiguating affective stimulus associations for robot perception and dialogue," in *Accepted in IEEE-RAS 18th International Conference on Humanoid Robotics (Humanoids)*, 2018.

[16] R. W. Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[17] S. Scheibe and L. L. Carstensen, "Emotional aging: Recent findings and future trends," *The Journals of Gerontology: Series B*, vol. 65, no. 2, pp. 135–144, 2010.

[18] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[19] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *2014 Canadian Conference on Computer and Robot Vision (CRV)*. IEEE, 2014, pp. 98–103.

[20] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[21] M. Gori, *Machine Learning: A Constraint-based Approach*. Morgan Kaufmann, 2017.

[22] P. Ekman, "The argument and evidence about universals in facial expressions," *Handbook of Social Psychophysiology*, pp. 143–164, 1989.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[24] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," *arXiv preprint arXiv:1803.05434*, 2018.