Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning

Srinivas Parthasarathy and Carlos Busso
Multimodal Signal Processing (MSP) Lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
sxpl20931@utdallas.edu, busso@utdallas.edu

Abstract
An appealing representation of emotions is the use of emotional attributes such as arousal (passive versus active), valence (negative versus positive) and dominance (weak versus strong). While previous studies have considered these dimensions as orthogonal descriptors to represent emotions, there are strong theoretical and practical evidences showing the interrelation between these emotional attributes. This observation suggests that predicting emotional attributes with a unified framework should outperform machine learning algorithms that separately predict each attribute. This study presents methods to jointly learn emotional attributes by exploiting their interdependencies. The framework relies on multi-task learning (MTL) implemented with deep neural networks (DNN) with shared hidden layers. The framework provides a principled approach to learn shared feature representations that maximize the performance of regression models. The results of within-corpus and cross-corpora evaluation show the benefits of MTL over single task learning (STL). MTL achieves gains on concordance correlation coefficient (CCC) as high as 4.7% for within-corpus evaluations, and 14.0% for cross-corpora evaluations. The visualization of the activations of the last hidden layers illustrates that MTL creates better feature representation. The best structure has shared layers followed by attribute-dependent layers, capturing better the relation between attributes.

Index Terms: emotion recognition, human-computer interaction, multi-task learning, deep neural networks

1. Introduction
Automatic emotion recognition systems can be broadly grouped into two tasks. The first group tries to identify discrete categories of emotions such as happiness, anger, and sadness [1, 2, 3]. The second group predicts values of emotional attributes such as arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong) [4, 5, 6]. Due to the complex nature of human interaction [7], it is hard to classify emotions into few simple distinct classes [8]. Extending the number of classes leads to sparseness and unbalance in the distribution of emotional classes. For these reasons, using few emotional attributes is highly appealing.

Most systems for predicting the values of emotional attributes are learnt from signals in either a single or multiple modalities. Studies have relied on acoustic, facial, and physiological signals such as electrocardiograph (ECG) and electroencephalogram (EEG) [9, 10]. The process usually consists of extracting high dimensional features to train classifiers that map these cues into each of these emotional attributes. There are certain disadvantages with systems trained this way.

• Most systems are trained to independently learn each emotional attribute (e.g., a valence regressor, an arousal predictor) [10, 11, 12]. These systems ignore the inherent dependency between emotional attributes. For example, previous studies have established dependencies between arousal and valence [13], and arousal and dominance [14, 15, 16].
• The input feature vector usually has high dimension [17]. The number of features is generally reduced using feature selection algorithm. The reduced feature set is then mapped into an emotional attribute. This process has to be separately repeated for each attribute, learning different feature representations.
• Previous studies have shown that the recognition of certain attributes is superior in some modalities than others [18, 19]. For example, valence values are better recognized using facial features than acoustic features [1, 20]. Similarly, arousal values are well recognized using acoustic features. When attributes are separately learned, we cannot exploit joint representations of features which might improve performance.

These observations suggest that an appealing solution to address these issues is to jointly learn multiple emotional attributes. Instead of considering emotional attributes as orthogonal descriptors, we should formulate machine learning algorithms that leverage the dependencies between them. These studies motivate us to build a unified framework to predict emotional attributes by leveraging their dependencies.

This study formulates the prediction of emotional attributes as a multi-task learning (MTL) problem. We consider arousal, valence and dominance. Using deep neural network (DNN) architectures, we train systems that learn to jointly model arousal, valence and dominance values. Using within-corpus and cross-corpora evaluations, we demonstrate that jointly learning emotional attributes leads to significant improvements over single task learning (STL), where each emotional attribute is separately modeled. In our framework, learning the attribute of interest is treated as the primary task and learning the other two attributes is treated as secondary tasks. This approach learns all three attributes, however, the network is optimized to increase the performance of the target attribute. This approach is repeated for arousal, valence and dominance where the attributes’ weights in the loss function are set over the development set. We restrict the evaluation to speech features, although the method is also appealing for multimodal features. The experimental evaluation shows gains in concordance correlation coefficient (CCC) across different experimental conditions when we use MTL over STL. We report gains up to 4.7% for within-corpus evaluations and 14% for cross-corpora evaluations. The visualization of the activation of the last hidden layers illustrates that MTL creates better feature representation. The best structure has shared layers followed by attribute-dependent layers, capturing the relation between attributes.

2. Related Work
Previous studies have shown the dependencies between emotional attributes. Lewis et al. [21] showed physiological evi-
dence for repeated patterns between arousal and valence. Olivia-

eria et al. [13] hypothesized and tested methods of integrating

predicted arousal and valence values by a weighted average op-
eration. Motivated by these studies, Nicolaou et al. [4] proposed

a hierarchical approach to estimate arousal and valence. The

first step independently predicts scores for arousal and valence.
The predicted scores are combined as features to derive the final

estimation for arousal and valence.

Few studies have focussed on joint learning multiple emo-
tional attributes. Xia and Liu [22] proposed a multi-task learn-
ing framework where the primary task was emotional categories
(e.g., happiness, anger), and the secondary task was either the

prediction of attribute scores or the classification of attribute

score into distinct classes (low, medium, high). They used deep belief networks (DBN) to train their multi-task problem.

Zhang et al. [23] proposed a multi-task framework with shared

hidden layers to jointly classify emotions using different emo-
tional representations (e.g., varied number of discrete classes,

quadrants in arousal-valence space). Their MTL framework

solved nine different representations, providing better perfor-

mance than separately solving each of them. Chang and Scherer

[24] recently proposed to jointly learn valence and arousal at-

tributes, where the main goal of the study was the use of deep

convolutional generative adversarial network to leverage unla-

beled data. This study is related to our work, but there are key differences between the studies. Chang and Scherer pro-

posed and tested multi-task models only for valence (as primary

task), since they hypothesized that only valence increases per-

formance by using joint representations with arousal. Furth-

ermore, they treated their problem as a three-class or five-class

classification problem, while our work uses regressors. Finally,

their results showed no improvements for multi-task learning of

valence along with arousal. The contributions of our work with

respect to previous studies are:

• Using MTL where the primary task is the target attribute

(e.g., arousal) and the secondary tasks are the prediction of the

other two attributes (e.g., valence, dominance)

• Exploring attribute-dependent layers on top of shared hidden

layers during MTL

• Extensive within-corpus and cross-corpora evaluations, dem-

onstrating the benefits of MTL over STL

3. Resources

3.1. Databases

This study uses the MSP-PODCAST corpus, which is be-
ing collected at The University of Texas at Dallas [25]. The
database is a collection of naturalistic, emotional data from au-
dio podcasts available on audio-sharing websites. These pod-
casts include discussion about a variety of topics including pol-
itics, movie review, science, technology, economics, business,
arts, culture, medicine, lifestyle and sports. The podcast ses-
dio podcasts available on audio-sharing websites. These pod-
cast segments are recorded by 135 speakers. We use segments from
50 speakers (5,024 speaking turns) as the test set, and segments
from 10 speakers (887 segments) as the development set. For
the within-corpus evaluation, the training set includes the rest
of the corpus (6,710 segments). This data partition attempts to
create speaker independent datasets for training, testing and de-
velopment sets.

We annotate the podcast segments for emotional content us-
ing a modified version of a crowdsourcing method introduced
by Burmania et al. [27]. At least five evaluators annotated emotional attributes using self-assessment manikins (SAMs):
 arousal (1 - very calm, 7 - very excited), valence (1 - very neg-

ative, 7 - very positive) and dominance (1 - very weak, 7 - very

strong). Although not used in this study, the segments are also
annotated for primary and secondary emotional classes [25].

For the cross-corpus evaluation, we use the recordings from
the USC-IEMOCAP [28] and the MSP-IMPROV [29] corpora.
Both databases were annotated in terms of arousal, valence and
dominance using SAMs. Further details can be found in the
papers describing the respective databases. For consistency, all
emotional attribute values are linearly scaled in the range [-1,1].

3.2. Acoustic Features

This study uses the popular feature set introduced for the com-
putational paralinguistics challenge in Interspeech 2013 [17].
The set uses a common approach to extract features for emotion
recognition. The approach extracts frame-by-frame low level descriptors (LLDs) such as fundamental frequency and Mel-
Frequency Cepstral Coefficients (MFCCs). For each speaking
turn, the approach extracts global statistics such as arithmetic
mean and standard deviation to the LLDs (e.g., mean of the en-
ergy). These global statistics are referred to as high level func-
tionals (HLF). The set contains 6,373 features extracted with
Opensmile [30].

4. Multi-Task Learning Framework

We formulate the prediction of emotional attributes as a regres-
sion problem that is solved with DNNs. The network takes the
acoustic features (Sec 3.2) as input and maps it into an attribute
score. The proposed approach predicts the value of an attribute
by jointly learning scores for arousal, valence and dominance.

If valence is the target attribute, the primary task is predicting
valence and the secondary tasks are predicting arousal and dom-
inance, where the corresponding weights are learned using the
validation set. Therefore, we still have three systems optimized
for arousal, valence and dominance.

We evaluate two MTL frameworks presented in Figures 1
and 2, where their difference is in the way the hidden layers are

Figure 1: DNN architecture for MTL. Hidden layer nodes are
shared by all attribute dimensions.
connected. The first framework, referred to as MTL1, shares all the nodes in the hidden layers between the three attributes. Figure 1 shows the DNN architecture for MTL1, which is the conventional approach in MTL. The second framework, referred to as MTL2, has shared nodes only in the first layer. These shared nodes create a joint feature representation for arousal, valence and dominance. For the second layer, however, the nodes are separately connected for each attribute. The attribute-dependent layers can learn representations that are optimized for each attribute. Figure 2 shows the DNN architecture for MTL2. Note that both frameworks predict an estimation for each attribute. However, we only consider the predicted value corresponding to the target emotional attribute.

We train the MTL frameworks by minimizing the mean square error (MSE). The approach will generate three different loss functions, one for each attributes (i.e., $L_{aro}$, $L_{val}$, $L_{dom}$). The overall loss function ($L_{ov}$) is a weighted sum of the three individual square losses. The weights are constrained by the parameters $\alpha$ and $\beta$, respectively. Equation 1 gives the overall loss for the MTL frameworks:

$$L_{ov} = \alpha \times L_{aro} + \beta \times L_{val} + (1 - \alpha - \beta) \times L_{dom}$$ (1)

where the values for $\alpha$ and $\beta$ vary between 0 and 1 in steps of 0.1 such that $\alpha + \beta \leq 1$. We evaluate the models on the validation set for all values of $\alpha$ and $\beta$. While we jointly learn the three attributes, we still maximize performance for each individual attribute (three different systems). Therefore, we separately choose the best parameters for each emotional attribute. For example, for arousal the best parameters may be $[\alpha = 0.6, \beta = 0.2]$, while for valence the best parameters may be $[\alpha = 0.2, \beta = 0.4]$. The best values for $\alpha$ and $\beta$ in the development set are then evaluated on the test set.

5. Experimental Evaluations

We conduct within-corpus and cross-corpora evaluations, which have different train sets. However, the development and test sets are consistent across the study (i.e., same regression task). We estimate whether the differences in performance for MTL and STL systems are statistically significant using a one-tailed z-test on difference in population proportions, asserting significance at $p$-value = 0.05. We compare MTL and STL systems with equivalent number of nodes.

5.1. Baseline Systems: Single Task Learning

We develop a STL baseline system that is trained to individually predict the value of arousal, valence and dominance (i.e., three different independent systems). The loss function is the MSE between the labels and predicted values of the emotional attributes. Notice that this STL can be formulated using MTL1 with $[\alpha = 1, \beta = 0]$ for arousal, $[\beta = 1, \alpha = 0]$ for valence and $[\alpha = 0, \beta = 0]$ for dominance.

5.2. Implementation

We evaluate the models using the *concordance correlation coefficient* (CCC). CCC measures the agreement between two variables, in our case, the true and predicted emotional attribute values

$$\rho_c = \frac{2\sigma_x \sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$ (2)

where $\mu_x$ and $\mu_y$ are the means of the two variables, $\sigma_x^2$ and $\sigma_y^2$ the variances of the two variables, and $\rho$ is their Pearson’s correlation. CCC has been used as the evaluation metric for several tasks including the AVEC2016 challenge on depression, mood and emotion recognition [9].

All networks take a vector of 6,373 acoustic features as input (Sec. 3.2). We standardize the features by subtracting the mean and dividing by the standard deviation across all training samples. We observed that few sentences contain unreliable features whose deviation from the mean is large. After standardization, we diminish the effect of unreliable features by attenuating (equating to zero) all features whose deviation from the mean is greater than 3. Notice that 99% of the samples fall within a deviation of 3 after mean-variance normalization. The subsequent layers of our neural network are tuned to predict the emotional attributes with unreliable feature values tuned to zero.

We evaluate various configuration for MTL1, MTL2 and the STL. We implement all the networks with two hidden layers on top of the input layer. We report performance when the number of nodes in each hidden layer is set to 256, 512 or 1024. The three attribute-dependent hidden layers in MTL2 have the same numbers of nodes as the shared hidden layers. We use rectified linear unit (ReLU) as the activation of the hidden layers. To increase the generalization of the networks and to avoid overfitting, we use dropout with a probability of 0.5 at the input layer and the first hidden layer. We use a linear layer as the output layer. We learn using stochastic gradient descent, with a learning rate of $1 \times 10^{-4}$ per sample, a momentum constant of 0.9, and a mini-batch size of 256.

5.3. Results on Within-Corpus Evaluations

For each condition (i.e., system, number of nodes per layer, target emotional attribute), we optimize $\alpha$ and $\beta$ to maximize CCC using the validation set. Figure 3 shows the surface illustrating the changes in CCC as a function of $\alpha$ and $\beta$. This figure corresponds to the MTL2 system for arousal with 512 nodes per hidden layers. The best performance is $\rho_{CCC} = 0.7789$, for $\alpha = 0.7$ and $\beta = 0.3$ (i.e., the arousal loss function has higher weights, as expected). The value for STL for arousal corresponds to the point $[\alpha = 1, \beta = 0]$, which is highlighted in the plot. The figure illustrates the gains achieved by joint learning the emotional attributes.

Table 1 shows the results for within-corpus evaluations where the train set has 6,710 segments from the MSP-PODCAST corpus. The table shows that MTL systems perform better than STL systems for almost all conditions (emotional attribute, number of nodes per layers). MTL2 achieves significantly better results than STL for valence and dominance. For

![Figure 2: DNN architecture for MTL2. The first hidden layer nodes are shared by all emotional dimensions. The second hidden layer is trained independently for each emotional attribute.](image-url)
arousal, results are significant only when we have 1024 nodes. The best framework is MTL2, which has a shared layer and attribute-dependent layers. This structure allows the network to learn the variation between the emotional attributes while jointly learning common representations.

5.4. Results on Cross-Corpora Evaluations

Table 2 show the results for cross-corpora evaluations. All speaking turns from the IEMOCAP and MSP-IMPROV corpora are used to train the models. With cross-corpora evaluations, the performance decreases due to the train and test mismatch. However, the gains in performance by using MTL over STL are more clear. MTL systems perform better than or equal to STL in almost all cases. Results are significantly better in most cases. The gains for dominance are especially high, reaching improvements up to 14%.

5.5. Visualization of Feature Representation

Finally, we visualize the activations of the final hidden layer of STL and MTL2 systems. We aim to illustrate the representations learnt by these methods. We rely on the t-SNE technique introduced by Maaten and Hinton [31], where we reduce the dimension of the activations to two using the default parameters. As an example, we consider the cross-corpora evaluation of arousal with 1024 nodes. Figures 4(a) and 4(b) show the 2-D feature representation for STL and MTL, respectively. For visualization, we split the ground-truth labels into three balanced clusters corresponding to red (arousal \(\in [-1, 0]\)), blue (arousal \(\in [0, 0.4]\)), and green (arousal \(\in [0.4, 1]\)). Although there is no clear separation between the clusters in both figures, we observe that clusters are better grouped with MTL2 (Fig. 4(b)). This result suggests that MTL2 learns better representations than STL. These figures correspond to preliminary experiments to understand feature representations for this task.

6. Conclusion

This study proposed a framework to jointly predict arousal, valence and dominance using multi-task learning. The target emotional attribute was considered as the primary task and the other emotional attributes were considered as the secondary tasks. The weights were learned from the development set to maximize the performance of the target attribute. The best performance was achieved with a structure that combines shared layers and attribute-dependent layers. This novel MTL structure learns shared representations across attributes, but allows each subtask to optimize the second layer to increase its performance. We show through within-corporus and cross-corpora evaluations that MTL achieves improvement in performance over STL. By visualizing the activation of the last hidden layers, we illustrates the better clustering provided by MTL.

For our future work, we want to generalize our results by extending evaluation to other corpora. The within-corpus evaluation showed improved performance as we increased the number of nodes per layers. We are currently collecting and annotating more speech samples, which will allow us to train more complex structures with more layers and nodes, leading to better performance. Furthermore, we want to use more sophisticated deep learning frameworks for this task such as recurrent neural networks and generative adversarial networks.
7. References


