Adversarial Attention Modeling for Multi-dimensional Emotion Regression

Suyang Zhu, Shoushan Li, Guodong Zhou*
Natural Language Processing Lab, School of Computer Science and Technology,
Soochow University, China
`syzhu@stu.suda.edu.cn, {lishoushan, gdzhou}@suda.edu.cn`

Abstract

In this paper, we propose an Adversarial Attention Network for the task of multi-dimensional emotion regression, which automatically rates multiple emotion dimension scores for an input text. Especially, to determine which words are valuable for a particular emotion dimension, an attention layer is learnt to weight the words in an input sequence. Furthermore, adversarial training is employed between two attention layers to learn better word weights via a discriminator. In particular, a shared attention layer is incorporated to learn public word weights between two emotion dimensions. Empirical evaluation on the EMOBANK corpus shows that our approach achieves notable improvements in r-values on both EMOBANK *Reader's* and *Writer's* multi-dimensional emotion regression tasks in all domains over the state-of-the-art baselines.

Sample Text:

I was very scared when the gunner started shooting the crowd. What a disaster!

Emotion dimension scores: *Valence* = 2.0, *Arousal* = 4.4, *Dominance* = 2.1

Figure 1: An example of multi-dimensional emotion regression. The dimensional emotion score ranges from 1.0 to 5.0. In this example, the word very in blue only suggests one emotion dimension (i.e, a high *Arousal* score). The word scared and disaster in red suggest two emotion dimensions. Specifically, scared suggests a low *Valence* score and a low *Dominance* score, while Disaster denotes a low *Valence* score and a high *Arousal* score.

1 Introduction

Emotion analysis aims to recognize human emotion expression in a given text (Mishne et al., 2005; Abdul-Mageed and Ungar, 2017). Typically, studies in emotion analysis can be divided into either emotion classification (Yang et al., 2007; Tripathi et al., 2017) or emotion regression (Yu et al., 2015; Wang et al., 2016a). While emotion classification aims to label an input text with a single or multiple emotion categories, emotion regression aims to rate a single or multiple emotion dimension scores of an input text through machine learning models. In this study, we focus on emotion regression.

Compared with enormous studies in emotion classification, studies in emotion regression have a late start much due to the inherent difficulty of the regression task and the lack of large-scale emotion regression corpora in high quality. Despite of its difficulty, emotion regression is more suitable for fine-grained emotion analysis and has gained an increasing attention recently due to the availability of several emotion regression corpora in the last few years (Preotiuc-Pietro et al., 2016; Yu et al., 2016; Hahn and Buechel, 2017). In principle, these emotion regression corpora apply the widely-admitted *Valence-Arousal* model or *Valence-Arousal-Dominance* model (Barrett, 2006) to describe emotions with a continuous real number space in two or three dimensions. Moreover, while different emotion classification corpora often apply different classification systems, they describe emotions with a limited number of discrete pre-defined emotion categories.

In the literature, most of the existing studies in emotion regression focus on a single emotion dimension by training multiple independent models for different emotion dimensions (Yu et al., 2015; Wang et al., 2016a). Hence in this paper, we seek to solve multi-dimensional emotion regression via a joint approach. Recently, attention mechanism

*Corresponding author

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 471–480
Florence, Italy, July 28 - August 2, 2019. ©2019 Association for Computational Linguistics

has been widely applied in sentiment and emotion classification (Wang et al., 2016b; Potamianos and Kokkinos, 2017). Likewise, in emotion regression, attention mechanism is supposed to be effective on determining what words are emotional for rating dimensional emotion scores. Figure 1 shows an example of emotion regression. Obviously, the dimensional emotion scores can be inferred from the colored words in this figure. Although the degree adverb, such as *very*, only suggests a high *Arousal* score, an emotional word often suggests more than one dimensional emotion score. This hints a possibility that the relationship between two emotion dimensions can be leveraged, which is overlooked by existing single-dimensional emotion regression studies.

In this paper, we try to model the multi-dimensional learning task as a multi-task learning task through adversarial learning. Recently, studies in multi-task learning via adversarial learning (Liu et al., 2017; Masumura et al., 2018), which tried to conduct adversarial learning (Goodfellow et al., 2014) between multiple tasks to learn task-specific features for achieving better performance for each task, has achieved a great success. We apply adversarial learning to model the task not only due to its capability of multi-task learning, but also due to its inherent collocability with attention mechanism. In the literature, adversarial learning has the difficulty in learning latent representations from discrete structures (e.g., sequence of word embeddings). Thus, most of existing studies in NLP apply adversarial learning with autoencoder-based models, which map a discrete word sequence into a continuous code space beforehand (Makhzani et al., 2015). In this study, we propose a more straightforward yet effective way to learn better representations via adversarial learning which directly learns continuous attention weights. This is done via an Adversarial Attention Network (AAN) which can leverage both advantages of adversarial learning and attention mechanism. AAN conducts adversarial learning between two attention layers to learn two sets of word weight parameters for two emotion dimensions. In this way, better weight information can be learned to represent words' importance for rating dimensional scores. Specifically, our proposed AAN has two features:

- First, AAN conducts adversarial learning between two attention layers to decide the values of words for rating two emotion dimension scores. In particular, we propose an adversarial training algorithm to learn two sets of better word weights which contribute to two emotion dimensions in two attention layers.

- Second, unlike existing single-dimensional emotion regression studies which separately train models for different emotion dimensions, AAN can leverage shared information between emotion dimensions (e.g., word *scare* contributes to both *Valence* and *Dominance* in the example shown in Figure 1) to better rate different emotion dimension scores, and thus achieve better regression results.

We apply AAN to the task of multi-dimensional emotion regression on a large-scale emotion regression corpus, namely EMOBANK, contributed by Hahn and Buechel (2017). Empirical evaluation on EMOBANK *Reader's* and *Writer's* multi-dimensional emotion regression tasks shows that AAN achieves significant improvements in r-values over several strong baselines. Furthermore, it also shows that adversarial training between two attention layers is more effective than simply applying attention mechanism individually to each emotion dimension, or simply training two regressors jointly for a pair of emotion dimensions.

2 Related Work

2.1 Emotion Regression

Compared with emotion classification, emotion regression had a late start due to the severe lack of large-scale annotated emotion regression corpora and the inherent difficulty of the regression task. Yu et al. (2015) implemented a lexicon-based weighted graph-based approach which models the relationship and similarity among emotion word nodes to rate the *Valence-Arousal* scores of emotion words. Their approach achieved the better performance over the simple linear regression approach, the kernel method, and the Pagerank algorithm. Preotiuc-Pietro et al. (2016) collected user information from Facebook, and built an English emotion regression corpus containing 2,895 texts. Wang et al. (2016a) proposed a regional CNN-LSTM-based approach to document-level emotion regression. Their approach first divided a whole text into several regions, and then extracted regional features from each region with multiple CNNs. By properly leveraging the fused regional features, an LSTM layer is finally applied to rating the *Valence-Arousal* scores of the whole text. Evaluation on several corpora showed the regional CNN-LSTM achieved a better performance over both the vanilla single-layered CNN and single-layered LSTM. Yu et al. (2016) constructed a Chinese emotion regression corpus, which contains 2,009 texts, from multiple online resources. Buechel and Hahn (2016) investigated mapping the dimensional emotion scores to an emotion category of a text. They first annotated the SemEval07: task 14 corpus with dimensional scores, and then constructed the mapping from dimensional emotion scores to the emotion categories by KNN. On the basis, Hahn and Buechel (2017) built an emotion regression corpus, namely EMOBANK, which contains over 10,000 texts.

2.2 Adversarial Learning

Due to the success of generative adversarial network (GAN) in image generation (Goodfellow et al., 2014), adversarial learning has drawn more and more attention in the recent years. In order to well address the instability issue in GAN's training, Arjovsky et al. (2017) proposed Wasserstein GAN (WGAN) to tackle the issue in GAN. Especially, WGAN applied the Wasserstein distance between two distributions instead of the JS divergence adopted in GAN to avoid the training instability issue due to the failure of the JS divergence to indicate the training process of the discriminator when there is few overlaps between two distributions.

In the recent years, NLP researchers began to apply adversarial learning to various NLP tasks. Zhang et al. (2016) and Zhao et al. (2017) constructed adversarial networks with CNNs and LSTMs to train text generation models. Wu et al. (2017) proposed two types of adversarial models which consist of CNNs and RNNs, respectively. They discussed the advantages and disadvantages of two implementations on two relation extraction datasets. Masumura et al. (2018) proposed an adversarial training approach for multi-task multilingual learning, which jointly conducts task discrimination among languages and language discrimination among tasks. Chen and Cardie (2018) applied adversarial learning to multilingual word representation learning which maps word embeddings in multiple languages to the same vector space.

In comparison, our study focuses on the task of multi-dimensional emotion regression. To the best of our knowledge, it is the first attempt which applies adversarial learning to emotion regression.

3 Adversarial Attention Network

In this section, we introduce AAN which conducts adversarial learning between a pair of emotion dimensions. Take the *Valence* dimension and the *Arousal* dimension as an example, Figure 2 illustrates the framework of the *Valence-Arousal* AAN. Besides the *Valence-Arousal* AAN, there are the *Valence-Dominance* AAN and the *Arousal-Dominance* AAN. Unless otherwise mentioned, in the rest of this section, we only introduce the detailed implementation of the *Valence-Arousal* AAN for convenience.

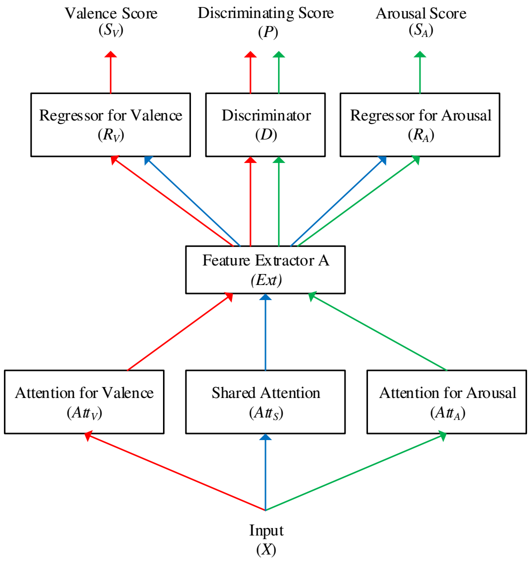

Figure 2: The framework of the *Valence-Arousal* Adversarial Attention Network which conducts adversarial learning between a pair of emotion dimensions. The frameworks of *Valence-Dominance* AAN and *Arousal-Dominance* AAN can be inferred in the same manner.

3.1 Attention Modeling

AAN takes a sequence of word vectors $X = [x_1 \ x_2 \ ... \ x_i \ ... \ x_k]$ of a text, which contains k words, as an input, where x_i denotes the word vector of the ith words in the text. The attention layer aims to learn a normalized weight vector

$A = [a_1\ a_2\ ...\ a_i\ ...\ a_k]$ from X by a one-layer LSTM to decide the value of a word vector, and finally output a weighted sequence:

$$\begin{aligned} X' &= Att(X) \\ &= diag(A)X \end{aligned} \tag{1}$$

$$A \ = softmax(LSTM(X)) \tag{2}$$

where Att denotes an attention layer, $diag(A)$ means to place the elements of A in the principal diagonal of a diagonal matrix with zero off-diagonal elements (Mulaik, 2009), X' denotes the weighted input sequence, and $softmax$ denotes the Softmax activation function for normalization. There are three attention layers, denoted as Att_V, Att_A, and Att_S, contained by an AAN. Att_V and Att_A decide which words are valuable for rating the *Valence* score and the *Arousal* score, respectively. Att_S is a shared attention layer to indicate which words contribute to the rating scores of both emotion dimensions:

$$X'_V \ = \ Att_V(X) \tag{3}$$

$$X'_A \ = \ Att_A(X) \tag{4}$$

$$X'_S \ = \ Att_S(X) \tag{5}$$

where X'_V, X'_A, and X'_S denote the weighted sequence returned by three attention layers, respectively.

3.2 Feature Extraction

The feature extractor of AAN (denoted as Ext) is trained to extract the feature vector from a weighted sequence returned by an attention layer. In this study, the feature extractor is implemented using a single-layered bidirectional LSTM ($BiLSTM$):

$$\begin{aligned} H &= BiLSTM(X') \\ &= [h_1\ h_2\ ...\ h_i\ ...\ h_k] \end{aligned} \tag{6}$$

In most of the previous studies, the hidden state of the last time step h_k from the output sequence H of BiLSTM layer is chosen as the feature vector. In this study, we further apply mean pooling to fetch richer textual information from the weight sequence:

$$\overline{h} \ = \ \frac{1}{k}\sum_{i=1}^{k} h_i \tag{7}$$

After mean pooling, h_k and $\overline{h}$ are concatenated as the output feature vector $Feat$ activated by the $tanh$ function:

$$\begin{aligned} Feat &= Ext(X') \\ &= tanh(h_k \oplus \overline{h}) \end{aligned} \tag{8}$$

where $\oplus$ denotes the concatenating operator. In AAN, the extraction of feature vectors from three weighted sequences is denoted as follows:

$$Feat_V \ = \ Ext(X'_V) \tag{9}$$

$$Feat_A \ = \ Ext(X'_A) \tag{10}$$

$$Feat_S \ = \ Ext(X'_S) \tag{11}$$

where $Feat_V$ and $Feat_A$ denote the features for *Valence* and *Arousal*, and $Feat_S$ denotes the shared feature which contributes to both emotion dimensions.

3.3 Dimensional Emotion Regression

The regressor rates an emotion dimension score. Since the regressor in AAN can be implemented in various ways as long as the gradients can be propagated in the network, to highlight the superiority of the proposed adversarial model, in this study, we implement the regressor simply with a single-layered full-connected neural network:

$$\begin{aligned} S &= R(Feat) \\ &= relu(W(Feat) + b) \end{aligned} \tag{12}$$

where S denotes the regression score of an emotion dimension, R denotes a regressor, W denotes the parameters of the full-connected layer, b denotes the bias term, $relu$ stands for the Relu activation function. In AAN, the *Valence* score S_V and the *Arousal* score S_A are denoted as follows. Note that the input of a regressor in AAN is the concatenation of the dimensional feature and the shared feature:

$$S_V \ = \ R_V(Feat_V \oplus Feat_S) \tag{13}$$

$$S_A \ = \ R_A(Feat_A \oplus Feat_S) \tag{14}$$

where R_V and R_A denote two regressors in AAN.

3.4 Emotion Dimension Discrimination

The discriminator D judges which emotion dimension an input feature vector contributes to. In the implementation of the D, we follow the work of WGAN, and apply the Wasserstein distance between two feature distributions as the loss function of the discriminator in order to provide a smoother measure for indicating the training process than KL divergence and JS divergence. In this study, the discriminator is implemented with a single-layered full-connected neural network to approximately fit the Wasserstein distance:

$$\begin{aligned} P &= D(Feat) \\ &= tanh(WFeat + b) \end{aligned} \tag{15}$$

where W denotes the parameters of the full-connected layer, b denotes the bias term, $tanh$ stands for the Tanh activation function. $P \in (-1, 1)$ stands for the discriminating result. In AAN, the closer the value of P is to 1, the more probably $Feat$ contributes to *Valence*. The discriminator outputs the results of $Feat_V$ and $Feat_A$:

$$P_V = D(Feat_V) \tag{16}$$

$$P_A = D(Feat_A) \tag{17}$$

where P_V and P_A denote the discriminating results of $Feat_V$ and $Feat_A$, respectively.

3.5 Adversarial Training

To adversarially train the model, we first train Att_V, Att_A, Att_S, R_V, R_A, and Ext by minimizing following regression losses. In this study, the mean square error is applied as the regression loss:

$$\min \frac{1}{n}\sum_{i=1}^{n}(S_{V_i} - T_{V_i})^2 \tag{18}$$

$$\min \frac{1}{n}\sum_{i=1}^{n}(S_{A_i} - T_{A_i})^2 \tag{19}$$

where S_{V_i} and S_{A_i} denote the regression scores of the *Valence* dimension and the *Arousal* dimension of the ith input sample, respectively. T_{V_i} and T_{A_i} denote the annotated true values of two emotion dimensions of the ith input sample. n denotes the total number of input samples.

Then, we update the parameters of D by maximizing the Wasserstein distance between two feature distributions:

$$\max \frac{1}{n}\sum_{i=1}^{n}(P_{V_i} - P_{A_i}) \tag{20}$$

where S_{V_i} and S_{A_i} denote the regression scores of two feature vectors extracted from the ith input sample. It is worthwhile to mention that we clip the parameters of D to a fixed absolute value at each training epoch. This training technique follows the research of Arjovsky et al. (2017) in order to meet the Lipschitz continuity which is required for using a full-connected layer to approximately fit the Wasserstein distance.

Finally, we update the parameters of Att_V and Att_A by adversarially fooling D:

$$\min \frac{1}{n}\sum_{i=1}^{n}(P_{V_i} - P_{A_i}) \tag{21}$$

Regarding the optimizing algorithm, in this study, we use different optimizers for different parts of our model. Att_V, Att_A, Att_S, and Ext apply Adam as their optimizers, while R_V, R_A, and D apply RMSProp as their optimizers. Parameters in the network are initialized with uniform samples in $[-\sqrt{6/(r+c)}, \sqrt{6/(r+c)}]$, where r and c are the numbers of rows and columns in the matrices (Glorot and Bengio, 2010).

4 Experimentation

In this section, we systematically evaluate our proposed AAN by applying it to the EMOBANK *Reader's* and *Writer's* multi-dimensional emotion regression compared with other baselines. For thorough evaluation, five-fold cross validation is applied in all experiments.

4.1 Experimental Settings

Dataset

In this study,the EMOBANK (Hahn and Buechel, 2017) is used in our experiments to evaluate the proposed approach. This multi-dimensional emotion regression corpus is available from the contributors' GitHub repository[1].

EMOBANK contains 10,548 texts annotated with 10,325 *Reader's* and 10,279 *Writer's* dimensional emotion scores, ranged from 1.0 to 5.0, in six domains. Table 1 gives the statistics of the numbers of texts in different domains on EMOBANK. In this study, we evaluate our approach in all the six domains of the EMOBANK corpus.

[1] https://github.com/JULIELab/EmoBank

Domain	*Reader's* Emotion	*Writer's* Emotion
News	2560	2540
Fictions	2824	2819
Blogs	1364	1349
Essays	1182	1238
Letters	1445	1383
Travel Guides	950	950
Total	**10325**	**10279**

Table 1: The distribution of annotated texts in each domain of the EMOBANK corpus. Note that the annotated texts of *Reader's* emotion and *Writer's* emotion are not exactly the same, which means not all the texts are annotated with both *Reader's* and *Writer's* emotion.

Hyper Parameters

Table 2 gives the most important hyper parameters of AAN. Note that AAN takes the sequence of word embeddings as its input. Here, the embedding look-up table is pre-trained with Word2vec, and is not dynamically updated during training.

Parameters	Value
word embedding dimension	300
feature dimension	150
learning rate (attention layer)	1e-4
learning rate (feature extractor)	8e-5
learning rate (regressor)	4e-5
learning rate (discriminator)	4e-5
batch size	64

Table 2: List of hyper parameters during AAN training. All the hyper parameters are tuned on a validating set randomly chosen from each domain of the EMOBANK corpus.

Evaluation Metrics

We apply the widely used Pearsons correlation coefficient r in all experiments as the evaluation metric for fair comparison because the contributors of EMOBANK also use r to evaluate the annotation quality between human annotators.

4.2 Baselines

In this study, the following baselines for emotion regression are implemented for fair comparison:

- **Deep CNN**: A CNN-based approach proposed by Bitvai and Cohn (2015). This approach applies multiple parallel CNNs to extract multiple n-gram features in a text, and is considered as one of the stat-of-the-art regression baselines for sentiment regression. In our implementation of Deep CNN, three parallel CNNs are applied to extract the unigram feature, the bi-gram feature, and the trigram feature in a text.

- **Regional CNN-LSTM**: A state-of-the-art emotion regression baseline proposed by Wang et al. (2016a). This approach first divides a whole text into several regions, and then extracts regional features from each region with multiple CNNs.

- **Context LSTM-CNN**: A state-of-the-art text classification baseline proposed by Song et al. (2018). This approach models the long-range dependencies within the classified sentences with an LSTM, and short-span features with a stacked CNN. We modified this approach by changing its activation function in order to return the dimensional emotion scores.

- **Attention Network**: A simpler counterpart of AAN. It contains only one attention layer, a feature extractor, and a regressor, for single-dimensional emotion regression.

- **Joint Learning**: Another simpler counterpart of AAN. It trains two regressors for two emotion dimensions in a joint learning style without any adversarial training technique. That is, this approach has the similar structure to AAN, except the absence of the discriminator. Here, three emotion dimension pairs are evaluated.

4.3 Experimental Results

Table 3 gives the performance of each approach in all six domains. Our proposed AAN notably performs better than other baselines, including the strong baseline Regional CNN-LSTM and Context LSTM-CNN in all cases. Furthermore, AAN outperforms its two counterparts (i.e., Attention Network and Joint Learning), justifying the effectiveness of the proposed adversarial learning approach. However, the overall r-values on EMOBANK are relatively low. This indicates the inherent difficulty of emotion regression on EMOBANK. As a reference, the average oracle r-value between human annotators of EMOBANK is about 0.6 (Hahn and Buechel, 2017).

Domain	Approach	*Reader's* Emotion			*Writer's* Emotion		
		V.	*A.*	*D.*	*V.*	*A.*	*D.*
News Domain	Deep CNN	0.288	0.150	0.136	0.217	0.060	0.127
	Regional CNN-LSTM	0.392	0.167	0.203	0.383	0.146	0.165
	Context LSTM-CNN	0.380	0.170	0.198	0.361	0.133	0.159
	Attention Network	0.349	0.167	0.194	0.351	0.135	0.158
	Joint Learning	0.377	0.169	0.200	0.366	0.139	0.161
	AAN	**0.424**	**0.187**	**0.238**	**0.414**	**0.175**	**0.179**
Fictions Domain	Deep CNN	0.228	0.201	0.157	0.187	0.170	0.164
	Regional CNN-LSTM	0.376	0.202	0.196	0.343	0.221	0.195
	Context LSTM-CNN	0.369	0.201	0.193	0.346	0.209	0.194
	Attention Network	0.355	0.198	0.190	0.331	0.208	0.190
	Joint Learning	0.371	0.202	0.195	0.333	0.214	0.194
	AAN	**0.405**	**0.209**	**0.218**	**0.384**	**0.243**	**0.204**
Blogs Domain	Deep CNN	0.256	0.281	0.118	0.220	0.249	0.131
	Regional CNN-LSTM	0.337	0.299	0.158	0.285	0.253	0.162
	Context LSTM-CNN	0.334	0.299	0.155	0.282	0.250	0.165
	Attention Network	0.325	0.291	0.149	0.280	0.242	0.159
	Joint Learning	0.330	0.287	0.154	0.282	0.249	0.160
	AAN	**0.353**	**0.308**	**0.165**	**0.299**	**0.260**	**0.171**
Essays Domain	Deep CNN	0.214	0.204	0.084	0.202	0.168	0.066
	Regional CNN-LSTM	0.334	0.241	0.081	0.303	0.179	0.059
	Context LSTM-CNN	0.320	0.239	0.077	0.299	0.169	0.064
	Attention Network	0.323	0.233	0.079	0.294	0.165	0.063
	Joint Learning	0.328	0.248	0.088	0.300	0.173	0.058
	AAN	**0.359**	**0.262**	**0.089**	**0.321**	**0.186**	**0.070**
Letters Domain	Deep CNN	0.316	0.283	0.194	0.257	0.207	0.222
	Regional CNN-LSTM	0.372	0.336	0.253	0.346	0.224	0.247
	Context LSTM-CNN	0.368	0.330	0.249	0.351	0.221	0.244
	Attention Network	0.358	0.322	0.239	0.331	0.211	0.239
	Joint Learning	0.364	0.329	0.245	0.350	0.218	0.243
	AAN	**0.380**	**0.351**	**0.265**	**0.378**	**0.254**	**0.261**
Travel Guides Domain	Deep CNN	0.202	0.161	0.155	0.196	0.188	0.106
	Regional CNN-LSTM	0.257	0.199	0.205	0.264	0.232	0.145
	Context LSTM-CNN	0.255	0.201	0.203	0.254	0.231	0.138
	Attention Network	0.248	0.189	0.196	0.251	0.217	0.132
	Joint Learning	0.251	0.202	0.202	0.255	0.224	0.130
	AAN	**0.267**	**0.216**	**0.226**	**0.277**	**0.240**	**0.151**

Table 3: The r-values of all the evaluated approaches to both *Reader's* and *Writer's* multi-dimensional emotion regression tasks on the EMOBANK corpus. Specifically, *V.*, *A.*, and *D.* are short for three emotion dimensions: *Valence*, *Arousal*, and *Dominance*, respectively.

In the *News* Domain, we can find that the performance in *Valence* is notably higher than those in other two dimensions. Moreover, the r-values in *Arousal* of *Writer's* emotion are lower than those of *Reader's* emotion. This indicates that a writer does not write a news article with too much emotional arousal in order to keep objectivity. For instance, the text *"Scam lures victims with free puppy offer."* relates to a negative emotion. This explains that the *Valence* scores of Reader's emotion and *Writer's* emotion are both low (<2.50). However, the *Arousal* score of *Reader's* emotion reaches 4.00, while the *Arousal* score of *Writer's* emotion is a medium of 3.25. This shows that in the *News* domain, the *Arousal* score of *Writer's* emotion tends to be a medium, even though a text can arouse a distinct *Reader's* emotion.

In the *Fictions* domain, the r-values in *Arousal* of *Reader's* emotion and *Writer's* emotion are much close compared with those in the *News* domain. This indicates that the writer of the *Fictions* domain writes texts with more distinct emotional arousal. For instance, the text *"She screamed: I havent socialized with Terras elite for most of my life!"* relates to a negative emotion, and the *Arousal* scores of *Readers* and *Writer's* emotion both reach 4.20. This shows that in the *Fictions* domain, a writer's emotional arousal is better represented by the *Arousal* score, and thus the r-value in *Arousal* of *Writer's* emotion is higher than that in the *News* domain.

Similar to the *Fictions* domain, the r-values in *Arousal* of *Reader's* emotion and *Writer's* emotion in the *Blogs* domain are very close. Furthermore, the r-values in *Arousal* are higher than those in the *Fictions* domain. This indicates that the emotion arousal in the *Blogs* domain is more distinct than that in the *Fictions* domain. For instance, the text *"lol Wonderful Simply Superb."* has extremely high score in *Valence* (4.8) and *Arousal* (4.8) of *Reader's* emotion, while its *Valence* and *Arousal* scores of *Writer's* emotion are also high (4.4 and 3.8, respectively). This implies that the writers of the *Blogs* domain express their emotion more frankly than those of the *Fictions* domain, and thus the regressor can better detect the emotion contained in the texts in the *Blogs* domain.

Unlike other domains, in the *Essays* domain, the r-values in *Dominance* of both *Reader's* emotion and *Writer's* emotion are extremely low. None of the baselines achieve an r-value in *Dominance* which is more than 0.1. The reason behind lies in that most texts in the *Essays* domain only objectively state realities. For instance, the text *"Moore's second hypothesis is that America's foreign policy may contribute to the belief that violence is an appropriate means to solve conflicts a hypothesis which is shared by many sociologists and psychologists."* only introduces the *"Moore's second hypothesis"* in an objective tone, while this kind of text is somehow hard to decide whether it expresses an active emotion or a passive emotion (i.e., whether the *Dominance* is high or low).

In the *Letters* domain, the performance in all dimensions reaches a high level in r-value compared with those in other domains. Specifically, there is no extremely low r-value (<0.20) in any dimension of either *Reader's* emotion or *Writer's* emotion. This implies that the writers of the *Letters* domain mostly write texts which relate to the real life of themselves or people around them. For instance, the text *"They do not have the resources necessary to purchase gifts or food for a holiday meal."* includes a pure emotion of writers, and such text can arouse more distinct emotion of readers.

Despite the overall lower performance than other domains due to the least text samples among all domains, there is no extremely low r-value achieved by any approach in the *Travel Guides* domain. Compared with the texts in the *Essays* domain, some texts in the *Travel Guides* domain state much about the histories and anecdota of the tourist attractions. However, besides the historical stories, for instance, the text *"Good for the health is just one of the many magical qualities that are attributed to these beautiful emerald-green or turquoise stones."* makes positive publicity for the tourist attraction in order to attract tourists, which contains a distinct positive emotion. Thus compared with the low r-values in *Dominance* in the *Essays* domain, the r-values in the *Travel Guides* domain are kept in a good level.

5 Conclusion

In this paper, we propose an Adversarial Attention Network (AAN) for multi-dimensional emotion regression. AAN takes the advantages from both adversarial learning and attention mechanism by conducting adversarial learning between two attention layers in order to learn better weighted information in a given text. Empirical evaluation on EMOBANK *Reader's* and *Writer's* three-dimensional emotion regression tasks shows the superiority of the proposed model with better performance over several state-of-the-art baselines. This indicates the effectiveness of the proposed adversarial learning approach to multi-dimensional emotion regression.

However, our proposed AAN still has several limitations. In our future work, we would like to improve the model structure and the adversarial learning algorithm. Moreover, we would like to seek a stable and controllable way to conduct adversarial learning among more than two objects. Last but not least, we would like to apply our approach to other heterogeneous texts-concerned NLP tasks.

Acknowledgement

This work is supported by three National Natural Science Foundation of China grants No.61702149, No.61672366, and No.61673290. We are very grateful to all reviewers for their meticulous and constructive comments and suggestions.

References

Muhammad Abdul-Mageed and Lyle H. Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 718–728.

Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein GAN. *CoRR*, abs/1701.07875.

Lisa Feldman Barrett. 2006. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and social psychology review*, 10(1):20–46.

Zsolt Bitvai and Trevor Cohn. 2015. Non-linear text regression with a deep convolutional neural network. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 180–185.

Sven Buechel and Udo Hahn. 2016. Emotion analysis as a regression problem - dimensional models and their implications on emotion representation and metrical evaluation. In *ECAI 2016 - 22nd European Conference on Artificial Intelligence, 29 August-2 September 2016, The Hague, The Netherlands - Including Prestigious Applications of Artificial Intelligence (PAIS 2016)*, pages 1114–1122.

Xilun Chen and Claire Cardie. 2018. Unsupervised multilingual word embeddings. In *EMNLP*, pages 261–270. Association for Computational Linguistics.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *CoRR*, abs/1406.2661.

Udo Hahn and Sven Buechel. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 578–585.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1–10.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. 2015. Adversarial autoencoders. *CoRR*, abs/1511.05644.

Ryo Masumura, Yusuke Shinohara, Ryuichiro Higashinaka, and Yushi Aono. 2018. Adversarial training for multi-task and multi-lingual joint modeling of utterance intent classification. In *EMNLP*, pages 633–639. Association for Computational Linguistics.

Gilad Mishne et al. 2005. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 workshop on stylistic analysis of text for information access*, volume 19, pages 321–327.

Stanley A Mulaik. 2009. *Foundations of factor analysis*. Chapman and Hall/CRC.

Alexandros Potamianos and Filippos Kokkinos. 2017. Structural attention neural networks for improved sentiment analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 586–591.

Daniel Preotiuc-Pietro, H. Andrew Schwartz, Gregory J. Park, Johannes C. Eichstaedt, Margaret L. Kern, Lyle H. Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in facebook posts.

In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, WASSA@NAACL-HLT 2016, June 16, 2016, San Diego, California, USA*, pages 9–15.

Xingyi Song, Johann Petrak, and Angus Roberts. 2018. A deep neural network sentence level classification method with context information. In *EMNLP*, pages 900–904. Association for Computational Linguistics.

Samarth Tripathi, Shrinivas Acharya, Ranti Dev Sharma, Sudhanshu Mittal, and Samit Bhattacharya. 2017. Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4746–4752.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xue-Jie Zhang. 2016a. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016b. Attention-based LSTM for aspect-level sentiment classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 606–615.

Yi Wu, David Bamman, and Stuart J. Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1778–1783.

Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. 2007. Emotion classification using web blog corpora. In *2007 IEEE / WIC / ACM International Conference on Web Intelligence, WI 2007, 2-5 November 2007, Silicon Valley, CA, USA, Main Conference Proceedings*, pages 275–278.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xue-Jie Zhang. 2016. Building chinese affective resources in valence-arousal dimensions. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 540–545.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xue-Jie Zhang. 2015. Predicting valence-arousal ratings of words using a weighted graph method. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 788–793.

Yizhe Zhang, Zhe Gan, and Lawrence Carin. 2016. Generating text via adversarial training. In *NIPS workshop on Adversarial Training*, volume 21.

Junbo Jake Zhao, Yoon Kim, Kelly Zhang, Alexander M. Rush, and Yann LeCun. 2017. Adversarially regularized autoencoders for generating discrete structures. *CoRR*, abs/1706.04223.