Predicting Video Affect via Induced Affection in the Wild

Yi Ding
yding@cs.ucsb.edu
University of California Santa Barbara

Radha Kumaran
rkumaran@ucsb.edu
University of California Santa Barbara

Tianjiao Yang
tianjiao_yang@ucsb.edu
University of California Santa Barbara

Tobias Höllerer
holl@cs.ucsb.edu
University of California Santa Barbara

ABSTRACT

Curating large and high quality datasets for studying affect is a costly and time consuming process, especially when the labels are continuous. In this paper, we examine the potential to use *unlabeled* public reactions in the form of textual comments to aid in classifying video affect. We examine two popular datasets used for affect recognition and mine public reactions for these videos. We learn a representation of these reactions by using the video ratings as a weakly supervised signal. We show that our model can learn a fine-graind prediction of comment affect when given a video alone. Furthermore, we demonstrate how predicting the affective properties of a comment can be a potentially useful modality to use in multimodal affect modeling.

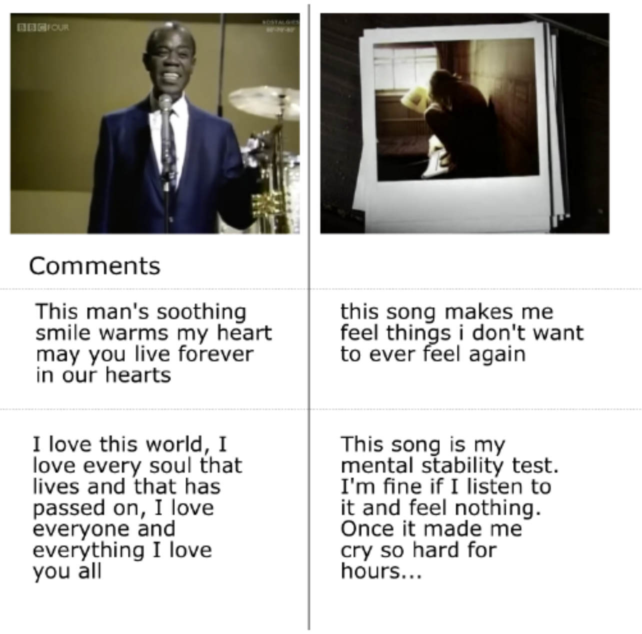
CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → **Machine learning approaches**;

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Yi Ding, Radha Kumaran, Tianjiao Yang, and Tobias Höllerer. 2020. Predicting Video Affect via Induced Affection in the Wild. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418838



Comments

This man's soothing smile warms my heart may you live forever in our hearts

this song makes me feel things i don't want to ever feel again

I love this world, I love every soul that lives and that has passed on, I love everyone and everything I love you all

This song is my mental stability test. I'm fine if I listen to it and feel nothing. Once it made me cry so hard for hours...

Figure 1: Example of comments for Youtube music videos. The left video has a low arousal and high valence rating. The right video has a low arousal and low valence rating. Corresponding comments from YouTube are shown below the figure and demonstrate correlation with the affect contents of the video.

1 INTRODUCTION

Affective computing uses computational techniques to model human psychophysiological states[32]. Researchers have tackled the recognition of these states uni- and multimodally. By recognizing these states, we can enable richer human-computer interaction by encoding human state beyond what is explicitly expressed. Application opportunities are broad and include the ability to automatically determine user opinion, empower individuals with better social cues, automated detection of misbehavior and many more.

Several high quality datasets have recently been developed to study this important problem [2, 19, 20]. These recent works provide both categorical measures and also incorporate continuous ratings of affect to capture subtle emotional differences. However, creating such high quality datasets is very resource intensive.

To alleviate this issue, we investigate whether we can use unlabeled public reactions to aid in determining the affect of a corresponding video. Since many emotional reactions are gathered in the lab by presenting a stimuli to induce an emotional response, then we would expect some similar reactions in the wild. For example if the average rating by test subjects in a lab for a particular video is 4 out of 5 for a happiness rating, then we should expect similar reactions and statistical measures by people in the public. Furthermore, we examine whether these responses can then be used to help determine the affect rating of the video.

We study these questions by examining comments found in the wild of videos used to induce emotions in a laboratory setting. We

ICMI '20, October 25–29, 2020, Virtual event, Netherlands
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7581-8/20/10.
https://doi.org/10.1145/3382507.3418838

attempt to learn a language model respective of the affect dimensions when given only the laboratory affect ratings of the video. The learned language features are then incorporated into a multimodal affect prediction model to determine video affect. Since each video has potentially millions of comments, designing an effective way to model this data can drastically reduce the video annotation burden.

Mathematically, our problem can be construed as learning a language representation over a mixture of Gaussians. We assume that each video occupies a region in affect space, and elicits emotional responses according to its distribution (an expected value as determined in a lab). We hypothesize that these emotional responses are reflected in the comments posted to a video. We take each video rating to induce an emotional response which can be used to "mold" the multiple comments associated with each video to region occupied by the video. That is, when given a weak prior in the form of expectations to video reactions, we would like to embed language in this affect space to fit its affective properties.

To learn this, we define a custom variational objective, an approach with demonstrated effectiveness in learning unsupervised sentence representations [6, 17]. By taking advantage of the smooth distribution learned by a VAE as well as weakly supervised information offered by the videos, we can mold the latent distribution of the comments such that they conform closer to the defined affect dimensions.

In summary, we provide the following contributions:

(1) We propose a novel problem for learning language representations from induced affect in the wild when given weakly supervised signals.
(2) We formalize the problem within the context of a Gaussian mixture and design an effective optimization method.
(3) We augment existing datasets with public reactions and make these augmentations publicly available.
(4) Experiments that indicate the potential to use induced signals in the wild for affect prediction tasks.

2 RELATED WORK

Affect representation has primarily been studied in two ways: as a categorical selection [9] or via dimensional representations such as the well known arousal-valence model [34, 38]. The research field of sentiment analysis often focuses on measures the valence dimension of affect: positive, negative or neutral [23]. We use the arousal-valence model, which measures affect on two orthogonal dimensions: arousal (the level of alertness/involvement) and valence (a pleasure-displeasure continuum) [34]. However, it should be noted that whether the dimensions are truly orthogonal can be deemed controversial [22].

Early works in affect computation primarily attempted to group people's emotional state into distinct categories. Over the years, researchers have expanded this capability to include more continuous affect representations [2, 20], by capture emotional state in addition to intensity which enable the modelling of modelling subtle difference. Multiple recent works have attempted to learn improved multimodal representations of affect-based data to improve downstream tasks such as video affect classification[41].

Emotion elicitation can by achieved by having subjects watch music videos [19, 26]. Visual [20] and audio stimuli [5] are among the most common modalities for inducing emotions. There has also been an increasing interest in using data collected in response to multimodal stimuli for the task of emotion recognition. Audio-visual stimuli have been studied in the form of monologues [44], conversations [33], and music videos [43]. The DEAP Database [19] contains records of EEG and peripheral physiological signals of participants who watched selected music videos, as well as the participants' self assessment of their emotional state after each trial. This has been used in emotion recognition and classification tasks [24] [39]. We use the arousal and valence values that have been provided for each video in the database as weak signals to supervise the learned representations in our model.

Language models seek to learn a representation and have been studied actively [3] [28] [27]. More recently, work has been done on using additional modalities in language modelling [18], incorporating symbolic knowledge into language models to allow for generation of rare words [1], and learning generalized representations of data for use in multiple language understanding tasks [25]. Frameworks to learn sentence representations by unsupervised learning methods have also been widely studied, such as asymmetric encoder-decoder structures[37], improvements to the VAE that learn semantics better [47], and the use of discourse relations to learn accurate sentence representations [29] [14]. Our approach, however, is to use a weakly supervised learning method, to embed language with explainable dimensions.

Some techniques used for affect recognition tasks include transfer learning [21] [8], attention modelling [49], and Tree-LSTMs [45]. Ghosh et al. [10] extend the LSTM model for text generation in conversations, allowing for control of the emotional content of the sentences generated. Another approach to controlling the emotion of generated sentences assume that the emojis in Twitter messages indicate the emotion of the conversation, and accordingly generates responses with appropriate emotion [48]. Song et al. [35] have explored affect-based text generation using not only explicitly emotional words, but also neutral words which express an emotion when combined in a specific pattern.

Traditional variational autoencoders (VAEs) usually incorporate a single Gaussian for regularizing latent variables, and Gaussian for the output as well. The output of VAE has been extended with mixture models and it has performed in unsupervised clustering [15] where the clusters are modelled by GMM, and [46] combine GMM and an uniform distribution to model major clusters and the remaining data, respectively. [16] adopt mixture models in the latent space for semi-supervised learning in classification problems, where different mixture components share parameters.

3 PROBLEM FORMULATION

We formalize the problem as follows: We denote V as a given set of videos that have been assigned an emotional rating, where a rating is a two-dimensional vector consisting of valence and arousal scores. Our task is to learn a mapping of comments $f : C \rightarrow \mathbb{R}^2$ such that $\mu_c := f(c)$ and μ_c reflects the true valence and arousal of the comments.

Each video has multiple ratings, and the mean $\mu_v \in \mathbb{R}^2$ and variance $\Sigma_v \in \mathbb{R}^{2\times2}$ of the ratings are given. When the number of raters is large enough, we can reasonably assume that all reactions towards a video v follows a normal distribution $\mathcal{N}(\mu_v, \Sigma_v)$. Explicitly, μ_v is the mean valence $\mu_{(1)}$ and mean arousal score $\mu_{(2)}$, while the covariance matrix Σ_v is diagonal, since we assume that valence and arousal are two uncorrelated [34], orthogonal criterion, i.e.,

$$\mu = (\mu_{(1)}, \mu_{(2)})^T,$$

$$\Sigma = \begin{pmatrix} \sigma_{(1)} & 0 \\ 0 & \sigma_{(2)} \end{pmatrix}.$$

For simplicity, we use $\text{diag}\{\sigma_{(1)}, \sigma_{(2)}\}$ to denote covariance.

The set of comments is denoted by C and comments associated with the video v are represented by C_v. Each video is given a rating μ_v which codes its effect on viewers. We are interested in exploiting the potential emotional influence of the video on any commenters. That is, the learned distribution of comments for a particular video should occupy the region described by the mean and variance of the video. Intuitively, while there may be a few deviating comments, a large proportion of the comments C_v in a video should agree roughly with the rating. Once we obtain the learned language model, we can then use the average comment scores for affect as an indicator for video affect.

4 LEARNING AN AFFECT EMBEDDING

A variational autoencoder (VAE) is an unsupervised architecture with demonstrated ability to produce quality representations of text [6]. They work by optimizing the parameter θ and maximizing the probability of each c such that:

$$P(c) = \int_{\mathcal{Z}} P_\theta(c|z)P_\theta(z)dz,$$

where $z \in \mathcal{Z}$ is a latent variable sampled by another function $Q(z|c)$ in order to reproduce c. VAEs are an extension of the standard autoencoder which imposes a prior distribution on z. It assumes that samples of z can be first drawn from a standard Gaussian distribution $p(z) \sim \mathcal{N}(0, I)$, where I is the identity matrix of the same column dimension with z. This has been empirically shown to learn smooth regions and enable better continuity in language representations. [6]

Hence we expect the distance between $Q_\phi(z|c)$ and $P_\theta(z|c)$ to be small. Mathematically, the standard VAE objective can be defined as [7] :

$$L_{\theta;\phi}(c) = E_{q_\phi(z|c)}[\log p_\theta(c|z)] - D_{KL}(q_\phi(z|c)||p(z)) \quad (1)$$

However, since the posterior distribution learned by the VAE is arbitrary, we cannot guarantee that a representation is learned in the dimensions that we want. Here we propose a simple tweak to use the valence and arousal ratings as a prior to shape the distribution. As a result, the ratings of comments from a video can be modelled as a two-dimensional uncorrelated Gaussian distribution $\mathcal{N}(\mu_p, \Sigma_p)$, where $\mu_p = (\mu_p^{(1)}, \mu_p^{(2)})$, $\Sigma_p = \text{diag}\{\sigma_p^{(1)}, \sigma_p^{(2)}\}$.

Giving us the following KL term instead of the KL term in (1) by the property of a diagonal matrix:

$$\begin{aligned} D_{KL}(Q||P) = \frac{1}{2}(&\text{tr}(\log\Sigma_p) - \text{tr}(\log\Sigma_q) - n \\ &+ tr(\Sigma_p^{-1}\Sigma_q) \\ &+ (\mu_p - \mu_q)^T\Sigma_p^{-1}(\mu_p - \mu_q)), \end{aligned}$$

where $\log\Sigma_p := \text{diag}\{\log\sigma_p^{(1)}, \log\sigma_p^{(2)}\}$.

4.1 Centered VAE (C-VAE)

Since it is known that stronger stimuli tends to produce a stronger emotional response, we introduce a second KL divergence term. To explain the reasoning, we introduce the definition of *uncertain response*: a response without a specific appropriate stimulus class [13]. Since stimuli with an uncertain response – close to the origin $(0, 0)$ – do not provide additional information regarding the prior, these videos should still contain comments that can vary wildly depending on personal preference which could "cover" the latent space.

Since the comments should match the center of the ratings with weight $1 - \lambda$, we construct the second KL term with the prior $p(x)$ sampled from the distribution of the entire dataset, i.e., for all $v \in V$. $p(x) \sim \mathcal{N}(\mu_x, \Sigma_x)$. For a total of N videos in V, $\mu_x = \frac{1}{N}\sum_{i=1}^{N}\mu_i$; and we compute $\sigma_x^{(r)}$, $r = 1, 2$ as follows:

$$\begin{aligned} (\sigma_x^{(r)})^2 &= E[(x^{(r)})^2] - (\mu_x^{(r)})^2 \\ &\approx \frac{1}{N}\sum_{i=1}^{N} E[(x_i^{(r)})^2] - (\mu_x^{(r)})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\{(\mu_i^{(r)})^2 + (\sigma_i^{(r)})^2 - (\mu_x^{(r)})^2\}. \end{aligned}$$

Using this mean and variance, we can create our new term $D_{KL}(Q||\mathcal{N}(\mu_x, \sigma_X))$. We use a λ term to weigh the potential variance in emotional responses based on its Euclidean distance to the original. This gives us our final loss function:

$$\begin{aligned} L_{\theta;\phi}(c) = &E_{q_\phi(z|c)}[\log p_\theta(c|z)] \\ &- \lambda D_{KL}(q_\phi(z|c)||p(z)) \\ &- (1-\lambda)D_{KL}(Q||\mathcal{N}(\mu_x, \Sigma_x)) \end{aligned} \quad (2)$$

4.2 Centered Gaussian Mixture VAE (CGM-VAE)

As it is a strong assumption that all comments to videos are normally distributed, we propose to use a Gaussian mixture prior. This allows us to be more accurate with respect to the center of the video ratings given by the prior $p(z)$. We extend the work of [11] which demonstrates numerous ways to approximate the KL divergence of Gaussian mixtures.

Recall that a Gaussian mixture consists of multiple Gaussian distributions and the proportion of each mixture component which is represented as a latent variable that yields the multinomial distribution. So we use unlabeled samples $\{y_i\}_{i=1,\dots,n}$ from n multi-dimensional Gaussians with known covariance matrices. This yeilds

Table 1: Summary statistics of dataset

	DEAP	MOSEI
# Videos	120	3228
# Videos with comments	82	764
# Comments	31481	17217

the mixture:

$$f_\theta(y) = \sum_{i=1}^{n} \pi_i \phi(y; \mu_i, \sigma_i^2 I_d)$$

where π_k is the mixing proportion of the k-th Gaussian distribution satisfying $\sum_{i=1}^{n} \pi_i = 1$, and $\phi(\cdot; \mu, \Sigma)$ denotes the density of a $\mathcal{N}(\mu, \Sigma)$ random vector in $\mathbb{R}^d$:

$$\phi(w; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(w-\mu)^T \Sigma^{-1}(w-\mu)\}$$

Specifically, for the total $n = N$ number of videos, we give equal weight to each part of the mixture, i.e., $\pi_k = \frac{1}{N}$. Since we don't distinguish the weight of each Gaussian distribution if there is no further information on the importance of the videos, the dimension of each Gaussian is $d = 2$ for the valence and arousal ratings.

Thus the second KL divergence follows

$$\begin{aligned} D_{KL}(Q||P) = & -\frac{1}{2}\mathrm{tr}(\log \Sigma_q) - 1 \\ & - \frac{1}{n}\mathbb{E}_Q\left[\log \sum_{i=1}^{N} \frac{1}{|\Sigma_i|^{\frac{1}{2}}} e^{-\frac{1}{2}A_k} \right], \end{aligned} \tag{3}$$

where $A_k := (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)$.

One possible way to make (3) computationally tractable is through Monte Carlo sampling. We draw K i.i.d samples $\{x_j\}_{j=1}^{n}$ from the distribution of Q, $\mathcal{N}(\mu_q, \Sigma_q)$:

$$\begin{aligned} D_{KL}(Q||P) = & -\frac{1}{2}\mathrm{tr}(\log \Sigma_q) - 1 \\ & + \frac{1}{2N} \sum_{k=1}^{n} \Big\{ \mathrm{tr}\big(\log \Sigma_k\big) + \mathrm{tr}\big(\Sigma_k^{-1} \mu_k \mu_k^T\big) \\ & + \frac{1}{K} \sum_{j=1}^{K} \mathrm{tr}\big(\Sigma_k^{-1} x_j (x_j^T - 2\mu_k^T)\big) \Big\}, \end{aligned}$$

where $A_k^j := (x_j - \mu_k)^T \Sigma_k^{-1} (x_j - \mu_k)$.

5 EXPERIMENT SETUP

We conduct two primary experiments to validate our optimization methods presented in equations 2 and 3. We examine the predictive power of our technique and its ability to learn embeddings through induced emotion signals. We then compare the fusion of our learned embeddings into a multimodal model to examine results compared on a state of the art benchmark. Our optimization methods are referenced as VAE (the standard VAE objective but with a modified prior), C-VAE (Centered VAE), and CGM-VAE (Centered Gaussian Mixture VAE).

5.1 Data

We apply our approach to use text comments on videos as a signal provider for instilled affect to augment two different datasets for affect prediction:

DEAP [19] provides affect annotations for music videos available on Youtube. We used the online subjective annotations video list containing 120 Youtube videos each with 14 to 16 ratings. A 9-point rating for valence, arousal, and dominance were collected, although we only examine the valence and arousal dimensions. We ask readers to refer to the original paper for detailed analysis [19]. For our use case, we used DEAP's valence and arousal ratings to embed comment language in a 2-dimensional space.

MOSEI [2] is a large multimodal sentiment and emotional dataset containing 23453 segments of videos by 1000 distinct speakers. Each video is an opinion video clip which is annotated in segments by 14 expert judges. Sentiment annotations on a Likert scale from -3 to 3 and Ekman emotions are annotated on a Likert scale of [0,3] from no evident emotion to high presence of emotion. For our use case, MOSEI's emotional space provided an additional 6-dimensional embedding vector for each comment. Additionally, since no video rating was provided, we took the mean of all segment-level ratings for each video as the overall video rating as input to our model.

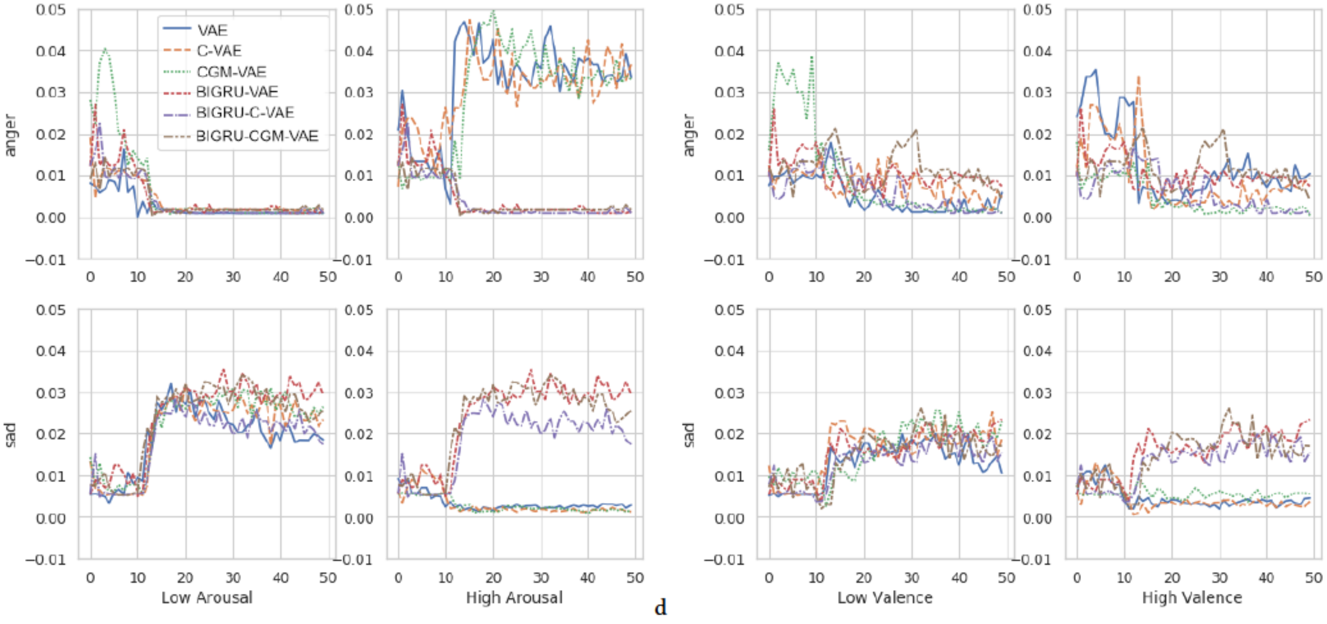
For all videos we crawled the available comments. The maximum number of comments per video was limited to 1000 and we exclude videos with no comments. Additionally, some videos were no longer available at the time of data collection. This resulted in 82 videos with comments for DEAP and 764 usable videos for MOSEI. Summary statistics are available in Table 1.

DEAP expresses instilled emotion (i.e. emotion of the viewer of a music video), while MOSEI characterizes the emotional state of the speaker in a video. User comments on a video may be more directly indicative of the user's emotion than the speaker's emotion, and we will evaluate this use case in Sections 6.1-6.3. Regarding the second case, it is our hypothesis that a causal connection sufficient for a distinctive signal likely exists as well, i.e., if I see a video of a happy/angry/sad person, I'm more likely to write a happy/angry/sad comment myself. We will show results on the MOSEI dataset in Section 6.4.

5.2 Preprocessing

The crawled comments are preprocessed to keep only the top-most level comment to remove any unrelated discussion using the `@user` expression. We also removed non-english comments and discarded sentences longer than 50 words, and shorter than 2 words for ease of language modeling. GloVE [31] word embeddings are used and kept fixed during training.

The dataset is split into an 80%-20% training-testing by randomly selecting 80% of the videos and their associated comments for training and the rest for testing. Validation is split from the training set during model tuning and for cross validation experiments in an 80%-20% fashion. For MOSEI evaluations, we observed the training, validation and testing splits provided by the MOSEI sentiment classification dataset.



(a) Arousal comparison over testing epochs epochs

(b) Valence comparison over testing epochs

Figure 2: Average of fine grained scores provided by LIWC for top and bottom 500 comments of learned embeddings. Anger is associated with high arousal but neutral valence, while sadness is associated with low arousal and low valence. A large inflection can be seen at approximately 12 epochs of training time due to the delay of kl-annealing. As can be seen, there is a correlation with some tested models

5.3 Network Architecture

We train our comments embedding network using a Recurrent Neural Networks (RNNs) connected in an end-to-end fashion [36] as the foundation for our modeling. We follow the the work from [6] closely in learning and optimization procedures, but use our learning objective.

Multiple network architectures were evaluated for our experiments. We used a Gated Recurrent Unit (GRU) as the base recurrent architecture. Single layer and 2 layer GRUs were used to evaluate our results. A 2-layer MLP is added to the output of the GRU to predict output distributions. Decoder architectures were varied with single, and 2-layer bidirectional GRU variants. The BiGRU tag is used to indicate the 2-layer bidirectional variant.

5.4 Hyperparameter Tuning

A the encoder and decoder hidden vector size was set to 100. Glove embeddings of 100 dimensions were used and kept fixed during model training. A two layer feed-forward network is attached to the output of the decoder GRU to predict word tokens. Monte Carlo samples used to approximate the gaussian mixture prior was set to 200. Although we experimented with different λ values, no large differences were noticed and were set at .5 for the entire experiment.

Hyperparameters were tuned on the validation set. The standard AdamW optimizer with all default options. A batch size of 128. Both Dropout and word dropout are used and is set to 0.2. Sigmoid KL annealing as used to train the evaluated VAEs offset by 15 epochs.

6 RESULTS & DISCUSSION

1) We provide empirical evaluation of the predicted video emotion ratings (Sections 6.1 & 6.2). 2) We show that our metric approximations agree with crowd-sourced user rated affect scores (Section 6.3). 3) We apply our technique to a large-scale public benchmark dataset for multimodal emotion analysis (MOSEI) and show that we can learn fine-grained emotion ratings for individual user comments while matching overall video emotions as the aggregate emotion of all user comments. We also demonstrate the ability for our embeddings to successfully extend an existing model with user comments as an additional dimension (Section 6.4).

6.1 Analysis on predictive power of comments

We perform an empirical evaluation of the learned language representations using the augmented DEAP dataset, as music performance is known for its ability to induce emotions, as demonstrated in lab studies. DEAP provides annotations for each video by multiple users in a valence, arousal and dominance space. We examine whether individual comments can be placed in a space close to their valence and arousal rating without the supervised information from the videos. We perform multiple experiments to correlate our predictions of valence and arousal scores with supervised tools which analyze language affect.

6.1.1 Supervised language analysis tools. The evaluations of the learned language representations are compared with two well-known tools for analyzing affect content in text. These tools are often used to provide distantly supervised information for related machine learning tasks[10] and are built on supervised knowledge. It is our expectation that if our learned representation *without* supervised information, demonstrates a positive correlation with existing supervised techniques, then we can expect that the video has 1) induced an expected emotion in the user and 2) our model can extract this information. This supervised information is *not* provided during training or testing time.

Two popular language analysis tools are used to analyze the video coments:

LIWC2015 [30] is a proprietary tool which produces scores for various dimension of language use. The typical output measures the fraction of words which fall under some variable.

The tone score is used to analyze our predicted valence score. It is a variable that measures the positive or negative tone of a text. Additionally, measures for anxiety, anger and sadness are provided which are typically associated with high, high and low arousal emotions [4].

VADER [12] is a lexicon and rule-based sentiment analysis tool designed for social media contexts. provides ratings for proportion of text which fall under categories of positive, neural, or negative as well as a normalized compound score. The compound score ranges from -1 to 1 and represents a summarizing of the overall positiveness or negativeness of the input text sequence. It is the expectation that the compound score most closely related to valence.

6.1.2 Metrics. As we do not have ground truth for sentence-level affect scores, we define an approximation based on expected emotion correlations:

$$V = \frac{S_c + S_t}{2} - .5$$
$$A = S_a + S_x - S_s$$

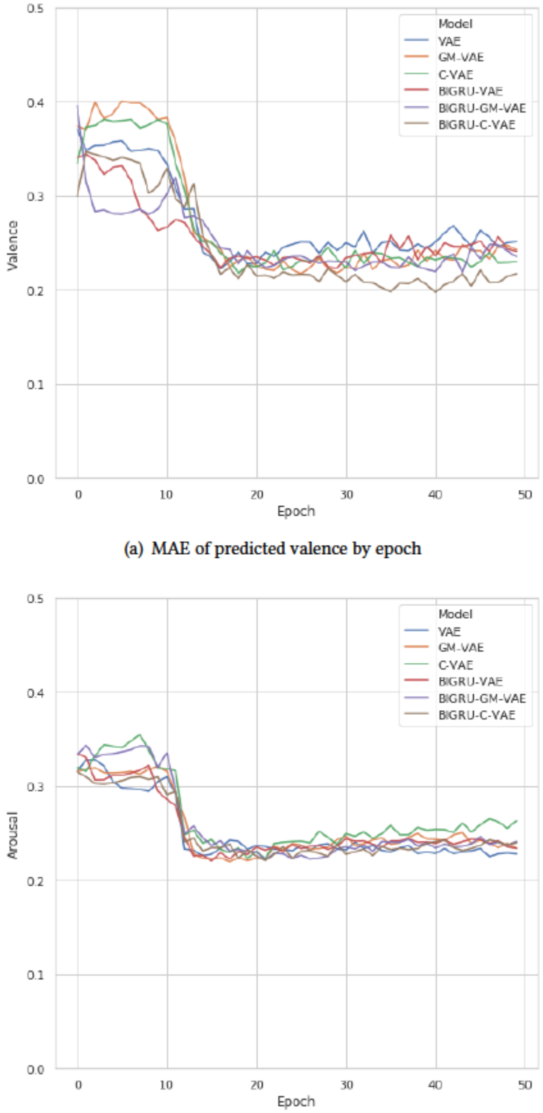
Where the S_c represents the normalized (between 0 and 1) in VADER compound score, and $S_t, S_a, S_x S_s$ represents the normalized in LIWC tone, anger, anxiety, and sadness scores respectively.

The tone and compound scores reflect the positive versus negative emotions present in the sentence. As there is no direct measure for arousal, we correlate with the LIWC measure for anger and sadness which are respectively positively and negatively correlated with arousal. We found that these metrics typically produced affect scores between -.5 and .5.

6.2 Video Affect Regression

We perform a 10-fold cross-validation evaluation with random initialization. The training set is split into 80% training and 20% validation. The MAE distance of predicted affect scores from our model with our valence and arousal metrics are shown in Figure 3. As can be seen, while the initial training shows large variations, all models eventually converge to a value closer to scores given by supervised approaches. Our Gaussian mixture optimization technique also shows the best average overall performance at epoch 50.

Table 2 shows the epochs with the minimum valence and arousal MAE values. As can be seen, the learned embeddings are moving away from a randomly embedded space into one which correlates with valence and arousal. Additionally, we see in figure 2, emotion scores from LIWC for each comment correlates with the expected embedding within valence and arousal space. We also see that comments are slowly conforming to the mold given by the prior distribution.



(a) MAE of predicted valence by epoch

(b) MAE of predicted arousal by epoch

Figure 3: 10-fold Cross-validation prediction of video rating with the DEAP dataset. MAE valence and arousal of our predicted comments ratings with defined metrics per epoch is shown.

6.3 Perception Study

We conducted a user study via Amazon Mechanical Turk asking users to rate the valence and arousal properties of learned comments representations. We compare the top and bottom ranked 100 comments for each dimension (valence and arousal) for each algorithm. All participating workers were from the US, with an

Table 2: Minimum MAE for valence and arousal ratings of video. Random shows the average MAE from randomly choosing scores and represents a baseline. Minimum possible value and best possible score is 0. As can be seen, our optimization method can learn embeddings that enable predictions close to the valence and arousal rating of the video.

Model	Valence MAE	Arousal MAE
Random	.33	.33
VAE	.222	.225
C-VAE	.217	.220
GM-VAE	.217	.220
BIGRU-VAE	.223	.221
BIGRU-C-VAE	.198	.221
BIGRU-GM-VAE	.218	.222

Table 3: User valence scores for the comments with model-estimated valence scores ranked in the top 100 and bottom 100. Scores range from 0 to 1. As we can see based on user ratings, the comments scored lower by our algorithm exhibit lower valence ratings and higher ranked comments received overall higher ratings from human raters.

Method	Bottom 100	Top 100
VAE	0.58 ± 0.02	0.60 ± 0.02
C-VAE	0.50 ± 0.03	0.67 ± 0.02
GM-VAE	0.56 ± 0.02	0.66 ± 0.02
BiGRU-VAE	0.54 ± 0.01	0.65 ± 0.02
BiGRU-CGM-VAE	0.54 ± 0.01	0.62 ± 0.02

Table 4: User arousal scores for the comments with model-estimated arousal scores ranked in the top 100 and bottom 100. Scores range from 0 to 1. A similar correlating trend is seen here with arousal scores.

Method	Bottom 100	Top 100
VAE	0.47 ± 0.02	0.61 ± 0.02
C-VAE	0.45 ± 0.02	0.57 ± 0.02
GM-VAE	0.50 ± 0.03	0.59 ± 0.02
BiGRU-VAE	0.51 ± 0.01	0.59 ± 0.02
BiGRU-CGM-VAE	0.51 ± 0.01	0.57 ± 0.01

approval rating greater than 98%. Workers provides ratings on a 5 point Likert scale, for valence as well as arousal of each comment.

Workers worked in batches of 16 comments with each sentence being rated by two unique workers. Inter-rater agreement was measured using Krippendorff's α as an ordinal metric, with $\alpha = 0$ representing perfect disagreement and $\alpha = 1$ representing perfect agreement. For this study $\alpha = 0.475$ for the *arousal* scale, and $\alpha = 0.686$ for the *valence* scale.

Table 3 shows the mean valence scores, as rated by the workers, of the top and bottom 100 comments. We perform a two population means t-test to compare the models, with a significance level $\alpha = 1\%$. C-VAE is significantly better than GM - VAE (p-value = 0.0444) and VAE (p-value = 0.0126) when identifying low-valence comments, and also outperforms VAE (p-value = 0.0135) in identifying high-valence comments.

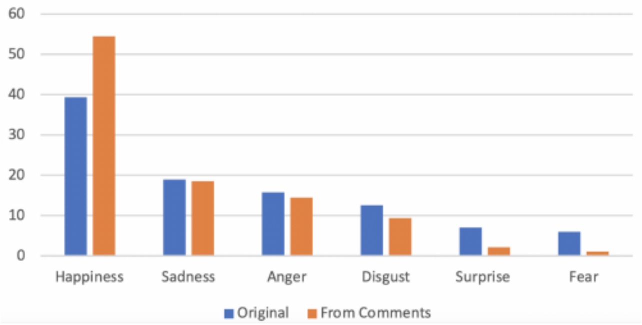


Figure 4: Percentage of emotions predicted by our learned comments compared to the original dataset.

The mean arousal scores are displayed in Table 4. The empirical analysis suggested that no model significantly outperforms another, and the results of this perception study indicate the same - no model performs significantly better than the others, both for low as well as high arousal sentences.

6.4 MOSEI benchmark

We evaluate the performance of our learned comment embeddings to aid in a multimodal affect classification task. We used the MOSEI sentiment unaligned dataset for this task. We compare against two state of the art techniques MulT [40] and Raven [42]. We follow the experiment setup from [40] and use their CTC augmentation of RAVEN.

We learn comment-level embeddings on the training set and accordingly predict the 6-dimensional emotion vector for each comment in the test set. Figure 4 shows that our video predictions capture the overall relative distribution of emotions from the original segment ratings.

Video-level emotional embeddings are generated by averaging the segment level predictions.

We concatenate the predicted video emotion ratings for each video onto the text embedding to fuse the comments context vector with segment-level text information. The resulting text representation is fed through the the network from [41] to perform the prediction.

Table 6 shows results from the MOSEI sentiment classification task which predicts sentiment classes for video segments. Our augmented affect predictions incurs but a slight effect on the final predicted segment sentiment scores. One limitation of our approach is that the user comments refer to the entire video, whereas MOSEI sentiment classification occurred on the level of shorter video segments.

Note also that the YouTube dataset makes up a small portion of the overall dataset and thus of the original 3228 videos, only 764 had comments (cf. Table 1). With additional comment information performance could potentially improve.

Table 5: Selected results, comparing the ground truth and predicted emotion of a video to that of a single comment on the same video. This demonstrates that the network learns better fine-grained embeddings than are provided by the overall video ratings. The six dimensions of emotion are happiness(H), sadness(Sa), anger(A), fear(F), disgust(D) and surprise(Su). A higher score indicates a stronger presence of the emotion.

Comment	GT Video Emotion						Predicted Video Emotion						Predicted Comment Emotion					
	H	Sa	A	F	D	Su	H	Sa	A	F	D	Su	H	Sa	A	F	D	Su
I liked lord of the flies in High School . It was really good i thought .	0.13	0.07	0.17	0.17	0.1	0	**0.18**	0.07	**0.07**	0.01	0.05	0	**0.3**	0.08	**0.04**	0	0.03	0.01
I hate how my English teacher just makes us write and give a public speech without even teaching how to do that ! ! ! ! ! ! !	0.33	0	0.17	0	0	0	**0.23**	**0.07**	0.06	0	**0.04**	0.01	**-0.05**	**0.09**	0.17	-0.02	**0.08**	0
Very interesting video, Melody, thank you!	0.17	0.17	0	0.06	0.06	0	**0.18**	**0.09**	0.08	0.01	0.05	0	**0.41**	**0.05**	0.04	0.01	0.02	0
As usual, she is excellent!	0.07	0.07	0.11	0.07	0.15	0	**0.17**	0.08	**0.06**	0	**0.04**	0	**0.31**	0.07	**0.02**	0	**0.02**	0
This made me realized how emotionally wounded I am, and I thought it was all normal.	0.19	0.1	0.05	0	0	0	**0.18**	**0.09**	0.06	0.01	**0.04**	0	**0.01**	**0.19**	0.06	0.01	**0.07**	0.02

Table 6: MOSEI Sentiment classification results on unaligned data. We see the augmentation of existing state of the art techniques improves its performance in a few situations.

Model	Acc7	Acc2	F1	MAE	Corr
CTC+RAVEN	45.5	75.4	75.7	0.664	0.599
MulT	50.1	81.0	81.2	.610	.681
MulT + C-VAE (ours)	49.1	81.2	81.5	.618	.681

Additionally, we provide a case study on the predicted emotion ratings of individual comments in table 5. As can be seen in multiple examples, despite the overall video not providing detailed emotion representations, we can still provide an effective prediction of the comment's emotions. For example, looking at the last row of table 5, we notice that the video has a positive emotion overall (H is larger than all the negative emotions). The comment however (which is clearly sad), is predicted to have Sa larger than H, which is accurate.

7 CONCLUSION

In this paper we examined the problem of learning sentence representations in affect space when given a weak prior in the form of a video affect rating. We introduced a novel problem and proposed and evaluated an effective optimization technique.

Our empirical evaluation of the predicted video emotion ratings show that it is possible to deduce affect from video content alone and that our approximation metrics agree with crowd-sourced user rated affect scores.

When applying our technique to a large-scale public benchmark dataset for multimodal emotion analysis (MOSEI), we show that we can learn fine-grained emotion ratings for individual user comments while matching overall video emotions as the aggregate emotion of all user comments. This demonstrates that our embeddings can successfully extend an existing multimodal model with user comments as an additional dimension. While we did not achieve the best performance here likely due to the differences in the way MOSEI and DEAP obtained affect labels.

Overall, we provided new augmentations of multi-modal video datasets and demonstrated the potential for reactive signals in the wild, in the form of user comments, to predict the affect induced by the videos, through modeling effective language representations in affect space.

In the future, we hope to explore better fusion of comment information, e.g. by aligning comments with references to specific portions of videos and other more fully understood content semantics.

ACKNOWLEDGMENTS

This work was supported in part by NSF award IIS-1845587 and ONR award N00014-19-1-2553. We thank the members of the Four Eyes Lab at UCSB for thoughtful discussions.

REFERENCES

[1] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2017. A Neural Knowledge Language Model. *arXiv:1608.00318 [cs]* (March 2017). arXiv:1608.00318 [cs]

[2] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.

[4] Jonah Berger. 2011. Arousal increases social transmission of information. *Psychological science* 22, 7 (2011), 891–893.

[5] Adnan Mehmood Bhatti, Muhammad Majid, Syed Muhammad Anwar, and Bilal Khan. 2016. Human emotion recognition and analysis in response to audio music using brain signals. *Computers in Human Behavior* 65 (2016), 267–275.

[6] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349* (2015).

[7] Carl Doersch. 2016. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908* (2016).

[8] Xin Dong and Gerard de Melo. 2018. A Helping Hand: Transfer Learning for Deep Sentiment Analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2524–2534. https://doi.org/10.18653/v1/P18-1235

[9] Paul Ekman and Dacher Keltner. 1997. Universal Facial Expressions of Emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture* (1997), 27–46.

[10] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. *arXiv:1704.06851 [cs]* (April 2017). arXiv:1704.06851 [cs]

[11] John R Hershey and Peder A Olsen. 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, Vol. 4. IEEE, IV–317.

[12] Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media.*

[13] Ray Hyman. 1953. Stimulus information as a determinant of reaction time. *Journal of experimental psychology* 45, 3 (1953), 188.

[14] Yacine Jernite, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557* (2017).

[15] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou. 2016. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148* (2016).

[16] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems.* 3581–3589.

[17] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[18] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *International conference on machine learning.* 595–603.

[19] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. Deap: A database for emotion analysis; using physiological signals. *IEEE transactions on affective computing* 3, 1 (2011), 18–31.

[20] Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. 2019. Deep Affect Prediction In-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. *International Journal of Computer Vision* 127, 6-7 (June 2019), 907–929. https://doi.org/10.1007/s11263-019-01158-4 arXiv:1804.10938

[21] Bernhard Kratzwald, Suzana Ilic, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Deep Learning for Affective Computing: Text-Based Emotion Recognition in Decision Support. *Decision Support Systems* 115 (Nov. 2018), 24–35. https://doi.org/10.1016/j.dss.2018.09.002 arXiv:1803.06397

[22] Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological bulletin* 139, 4 (2013), 917.

[23] Bing Liu. 2012. Sentiment Analysis and Opinion Mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[24] Wei Liu, Wei-Long Zheng, and Bao-Liang Lu. 2016. Emotion Recognition Using Multimodal Deep Learning. In *Neural Information Processing (Lecture Notes in Computer Science)*, Akira Hirose, Seiichi Ozawa, Kenji Doya, Kazushi Ikeda, Minho Lee, and Derong Liu (Eds.). Springer International Publishing, Cham, 521–529. https://doi.org/10.1007/978-3-319-46672-9_58

[25] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4487–4496. https://doi.org/10.18653/v1/P19-1441

[26] Maryanne Martin. 1990. On the Induction of Mood. *Clinical Psychology Review* 10, 6 (Jan. 1990), 669–697. https://doi.org/10.1016/0272-7358(90)90075-L

[27] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur. 2010. Recurrent Neural Network Based Language Model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*, Vol. 2. 1045–1048.

[28] Andriy Mnih and Geoffrey Hinton. 2007. Three New Graphical Models for Statistical Language Modelling. In *Proceedings of the 24th International Conference on Machine Learning - ICML '07*. ACM Press, Corvalis, Oregon, 641–648. https://doi.org/10.1145/1273496.1273577

[29] Allen Nie, Erin D Bennett, and Noah D Goodman. 2017. Dissent: Sentence representation learning from explicit discourse relations. *arXiv preprint arXiv:1710.04334* (2017).

[30] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015.* Technical Report.

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[32] Rosalind W Picard. 2000. *Affective computing.* MIT press.

[33] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 527–536. https://doi.org/10.18653/v1/P19-1050

[34] James Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39 (Dec. 1980), 1161–1178. https://doi.org/10.1037/h0077714

[35] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating Responses with a Specific Emotion in Dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3685–3695. https://doi.org/10.18653/v1/P19-1359

[36] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems.* 3104–3112.

[37] Shuai Tang, Hailin Jin, Chen Fang, Zhaowen Wang, and Virginia R de Sa. 2017. Speeding up context-based sentence representation learning with non-autoregressive convolutional decoding. *arXiv preprint arXiv:1710.10380* (2017).

[38] Robert E. Thayer. 1990. *The Biopsychology of Mood and Arousal.* Oxford University Press.

[39] 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 *Twenty-Ninth IAAI Conference.*

[40] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *arXiv preprint arXiv:1906.00295* (2019).

[41] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).

[42] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7216–7223.

[43] Ashkan Yazdani, Jong-Seok Lee, Jean-Marc Vesin, and Touradj Ebrahimi. 2012. Affect Recognition Based on Physiological Changes during the Watching of Music Videos. *ACM Transactions on Interactive Intelligent Systems* 2, 1 (March 2012), 1–26. https://doi.org/10.1145/2133366.2133373

[44] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems* 31, 6 (Nov. 2016), 82–88. https://doi.org/10.1109/MIS.2016.94

[45] Yuan Zhang and Yue Zhang. 2019. Tree Communication Models for Sentiment Analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.* 3518–3527.

[46] Qingyu Zhao, Nicolas Honnorat, Ehsan Adeli, Adolf Pfefferbaum, Edith V Sullivan, and Kilian M Pohl. 2019. Variational autoencoder with truncated mixture of gaussians for functional connectivity analysis. In *International Conference on Information Processing in Medical Imaging.* Springer, 867–879.

[47] Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. *arXiv preprint arXiv:1804.08069* (2018).

[48] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. *arXiv:1711.04090 [cs]* (May 2018). arXiv:1711.04090 [cs]

[49] Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial Attention Modeling for Multi-Dimensional Emotion Regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 471–480. https://doi.org/10.18653/v1/P19-1045