Predicting Microblog Sentiments via Weakly Supervised Multi-Modal Deep Learning

Fuhai Chen, Student Member, IEEE, Rongrong Ji, Senior Member, IEEE, Jinsong Su, Member, IEEE, Donglin Cao, Member, IEEE, and Yue Gao, Senior Member, IEEE

Abstract—Predicting sentiments of multi-modal microblogs composing of text, image, and emoticon has attracted ever-increasing research focus recently. The key challenge lies in the difficulty of collecting sufficient amount of training labels to train a discriminative model for multi-modal prediction. One potential solution is to exploit the labels collected from social media users, which is however restricted by negative effect of label noise. Besides, we have quantitatively found that sentiments in different modalities may be independent, which disables the usage of previous multi-modal sentiment analysis schemes in our problem. In this paper, we introduce a Weakly Supervised Multi-modal Deep Learning (WS-MDL) scheme towards robust and scalable sentiment prediction. WS-MDL learns convolutional neural networks iteratively and selectively from “weak” emoticon labels, which are cheaply available and noise containing. In particular, to filter out the label noise and to capture the modality dependency, a probabilistic graphical model is introduced to simultaneously learn discriminative multi-modal descriptors and infer the confidence of label noise. Extensive evaluations are conducted in a million-scale, real-world microblog sentiment dataset crawled from Sina Weibo. We have validated the merits of the proposed scheme by quantitatively showing its superior performance over several state-of-the-art and alternative approaches.

Index Terms—Sentiment prediction, Weakly supervised learning, Multi-modality, Deep learning

I. INTRODUCTION

Nowadays, we have witnessed a rapid spread of microblogs like Twitter and Sina Weibo in the Web. For instance, Sina Weibo hits a remarkable amount of 261 million monthly active users by March 2016. Such a rich social media repository provides an emerging channel for web users to express their sentiments. More interesting, such expression is more and more tending to a multi-modal way, which is composed of image, video, short text, and emoticons. Such multi-modal social media has posed significant application prospects, ranging from event monitoring, social network analytics, to commercial recommendations, etc. [1], [2], [3], [4], [5], [6], [7], [8]. However, for sentiment analysis, most existing works [9], [10], [11], [12], [13] still retain on analyzing the textual modality alone, while analyzing sentiments from visual and other modalities retains as an open problem [14], [15], [16], [17], [18].

It has been revealed in cognitive science that different modalities have their individual characteristics (semiotically, semantically and cognitively) in the sentiment perception of human being [19], which inspires us to analyze sentiments of microblog from a multi-modality aspect. Among the most pioneering works, Chen et al. [17] proposed a multi-modal hypergraph learning model for microblog sentiment prediction, which captures the correlation and independence among different modalities. However, the scalability, i.e., the ability of extending the schemes into large-scale applications, of [17] is restricted due to the computational complexity of hypergraph learning. For another work, You et al. [20] proposed a cross-modality regression model using deep neural networks, which considers the modality consistency to predict visual-textual sentiments accurately. However, such model is still not scalable due to the lack of sufficient labels.

To sum up, predicting multi-modal sentiments of microblogs retains as an open problem. The key challenges lie in the difficulty in learning discriminative representation across multiple modalities, as well as the limitation in collecting sufficient label. On one hand, due to the modality independence (i.e., each modality keeps its individual information under the task of the sentiment prediction) as stated in [17], on the aspect of feature fusion, it is difficult to find a unified feature representation across multiple modalities to learn a classifier.
In this paper, we propose a **Weakly-Supervised Multi-modal Deep Learning (WS-MDL)** model to learn from the cheaply-available noisy labels contributed by social network users, which handles the above two challenges in a unified framework. Fig.2 overviews the flowchart of the proposed framework (definitions of notations and detailed explanations will be described in Section III). In particular, we treat sentiment from the emoticon channel contributed from the social media users as weak labels to initialize the model learning. The subsequent task is two-fold, *i.e.* training multi-modal sentiment classifier from noisy labels, and infer label noise. To train multi-modal sentiment classifier, a multi-modal convolutional neural network is presented, which learns discriminative joint feature representation from different modalities (*a.k.a.*, CNN\(^1\) for visual and DCNN\(^2\) for textual modalities, respectively). To infer label noise, a weakly supervised learning paradigm is introduced, which formulates the correlation among the predicted labels in different modalities by a probabilistic graphical model. In such a manner, optimal parameters in this graphical model are estimated to simultaneously determine the contributions of intra-modal features (textual and visual features), and to infer the confidence of noisy labels.

To evaluate the effectiveness of the proposed model, we further release an 80K microblog sentiment dataset crawled from Sina Weibo\(^3\). We have conducted extensive experimental results and comparisons with the existing and state-of-the-art methods, including Cross-modality Logistic Regression (CBM-LR) [24], Cross-modality Logistic Regression (CBM-SVM) [24] and Hypergraph learning (HGL) [17]. The superior performance gains have demonstrated the merits of the proposed model.

**II. RELATED WORK**

**Visual Sentiment Prediction.** Most existing works of sentiment prediction mainly focus on analyzing the textual channel alone [9], [25], [10], [11], [23]. However, the need of visual sentiment prediction has been emerging recently. To predict visual sentiment, several recent works have been proposed, such as mid-level detectors [26], [27], [28], [29] and deep features [30], [31], [32], [33], [34]. For instance, Adjective Noun Pair (ANP) proposed by Borth et al. [26] is an explainable mid-level representation for visual sentiment prediction. For another instance, Yuan et al. [27] proposed the segmentate detectors to find diffusion patterns in social images. However, the existing works typically integrate sentiments predicted different modalities by direct fusion, which, as quantitatively shown subsequently in Fig.3, is highly independent and complicated. More importantly, existing works rely on “clean” labels to train robust classifiers to predict sentiments, which are hard to be extended to very large-scale applications.

---

\(^1\)CNN: Convolutional Neural Network [21]. It consists of 5 convolutional layers and 2 fully connected layers which has the pretrained model on ImageNet dataset [22].

\(^2\)DCNN: Dynamic Convolutional Neural Network [23]. It consists of 3 convolutional layers, a special dynamic k-max pooling layer, and a fully connected layer, which achieves the state-of-the-art performance for textual sentiment prediction.

\(^3\)http://weibo.com
Multi-modal Deep Learning for Sentiment Prediction. Multi-modal deep learning for sentiment prediction has attracted much research focus recently. For instance, You et al. [20] proposed a joint visual-textual model for sentiment analysis, which employs CNN and Distributed Paragraph Vector in feature extraction. Poria et al. [35] proposed a deep model to fuse speech, voice tone, and facial expressions for multi-modal emotion recognition and sentiment analysis. You et al. [36] proposed a new multi-modal deep learning framework by integrating textual and visual information in a structured fashion for joint visual-textual sentiment analysis. Poria et al. [37] proposed models based on a pre-trained convolutional neural network for extracting sentiment, emotion and personality features for sarcasm detection. Cambria et al. [38] propose a deep-learning-based framework for multi-modal sentiment analysis and emotion recognition by combining visual, text and audio features. Poria et al. [39] proposed to fuse audio-, visual- and textual-based affect detectors for multi-modal affect recognition. However, all aforementioned works are still limited to the need of a sufficient amount of “clean” labels.

Weakly Supervised Learning. To exploit massive noisy labels, weakly supervised learning is a popular solution. For instance, Lee [40] trained a network with both labeled and unlabeled data, which assigns labels to unlabeled data that has maximum predicted probabilities. Raykar et al. [41] proposed a Bayesian framework for supervised learning in the presence of contradicted labels. More recently, Oquab et al. [42] proposed a weakly supervised CNN to predict the location of objects in images by using only the image-level labels. Learning from noisy labels is also recently investigated for deep models. Mnih and Hinton [43] proposed two robust loss functions for DNN to deal with label noise. In [44], a novel convolutional network is proposed for improving the model robustness against both the label noise and the outlier noise, i.e., the noise from the randomly-selected data in other discrepant dataset.

III. WEAKLY SUPERVISED MULTI-MODAL DEEP MODEL

A. The Preliminary

We define clean label for a tweet as the sentiment label provided by expert annotators for a tweet. We define noisy label as the using of emoticons (if any) in a tweet. As quantitatively shown in [17], sentiments exhibit in different modalities could be diverse in a tweet, e.g., the sentiment of textual content appears to be positive while the visual one appears to be negative, which forms a sentiment situation. In this paper, the contribution weight of an emoticon for corresponding sentiment situation is defined by a sentiment contribution weight, termed SC-weight. As a preliminary study, Fig.3 shows the heatmaps of distance matrices among instances on the three modalities, (i.e., textual, emoticon and visual content). We quantitatively checked and validated that the emoticon modality is very consistent to the actual sentiment labels. However, due to its low frequency (only 32% tweets contain emoticon), it is hard to directly utilize emoticon for sentiment prediction. Instead, it is reasonable to assume the sentiments from emoticon channel are “weak” noisy labels to be input to the subsequent learning model.

Given a set of N multi-modal tweets \( D = \{<x_1, y_1, z_1>, ..., <x_N, y_N, z_N>\}\) (including text, image and emoticon) for training, where \(x_i\), \(y_i\) and \(y_i\) denote image, text, and emoticon of the i-th tweet, respectively. \(y_i^j \in \{0, 1\}\) denotes the occurrence of the j-th emoticon in the i-th tweet, where \(j \in \{1, ..., R\}\) (the j-th dimension) represents the j-th index in the emoticon list. We further define the visual sentiment label \(y_i\) and the textual sentiment label \(y_i\) as the latent variables, where \(y_i\) and \(y_i\) are both three-dimensional vectors corresponding to positive, neutral, and negative polarities, respectively. \(z_i\) denotes whether the polarities of image and the text are consistent, which is a latent variable to determine the correlation among both modalities. We then form K sentiment situations (\(K = 3 \times 3 \times 3\)), in which the k-th situation (\(k \in \{1, ..., K\}\)) is denoted as \(\langle y_i^1, y_i^2, y_i^3, z_i \rangle_k\). Fig.4 shows the graphical representation of the multi-modal noise modeling, in which the observable variable \(y_i\) is decided by \(y_i\) and \(z_i\), where multi-labels \(y_i\)
represent the combination of $\{\gamma_i^1, \ldots, \gamma_i^R\}$. Let $^v\theta$, $^t\theta$ and $\alpha$ be the parameters to be learned in our model, which denote visual CNN parameters, textual DCNN parameters and the SC-weight of emoticon, respectively. For clarity, we illustrate several important notations and their definitions throughout the paper in Tab. I.

During the iteration of model learning, we target at picking out incorrect “weak” labels by estimating their corresponding weights. Since there might be more than one emoticon labels in a tweet, we make an assumption that the emoticons in a microblog tweet are independent with each other. For the $i$-th tweet, given the $k$-th sentiment situation $\langle y_i^k, y_i, z_i \rangle$ and its corresponding SC-weight $\alpha_k$, we have the probability distribution of the $j$-th observed emoticon label $\gamma_i^j$, which is denoted as $p(\gamma_i^j | \langle y_i^k, y_i, z_i \rangle ; \alpha_k)$. Given the above conditional independence assumption, we have

$$p(\gamma_i^j | \langle y_i^k, y_i, z_i \rangle ; \alpha_k) = \prod_{j=1}^{R} \left( \frac{\alpha_k^{\gamma_i^j}}{1 - \alpha_k^{\gamma_i^j}} \right)^{\gamma_i^j},$$

where the combination of $\gamma_i^1, \ldots, \gamma_i^R$ is represented as $\gamma_i$ for the sake of simplicity.

### B. Weakly Supervised Learning Framework

In this section, we introduce the proposed weakly supervised multi-modal deep learning (WS-MDL). The detailed training procedure will be given in Section IV.

As shown in Fig. 2, the proposed WS-MDL model consists of 3 components, i.e., visual sentiment prediction, textual sentiment prediction, and multi-modal noise estimation. The learning is conducted by the following workflow:

1. Both visual (AlexNet [21]) and textual (DCNN [23]) deep models are initialized by using a limited number of “clean” labels.
2. We take the image and text of a tweet as the inputs of the corresponding deep model components to get their sentiment predictions;
3. The outputs of these two components treated as latent variables, together with the emoticon as the observed variables, are used to jointly optimize the CNN parameter $^v\theta$, DCNN parameter $^t\theta$, and SC-weight of emoticon $\alpha$ by Expectation-Maximization algorithm.

### C. Parameter Learning

We further make an assumption that the instances are independently sampled. Now given the observations $D_i$ of the $i$-th instance, the likelihood of parameter set $\vartheta = \{^v\theta, ^t\theta, \alpha\}$ can be written as:

$$p(D_i | \vartheta) = p(\gamma_i | x_i; \vartheta),$$

where the $i$-th sample is denoted as $x_i = \{^v x_i, ^t x_i\}$. Further, we expand the likelihood function with combinations of latent variables $\langle y_i^k, y_i, z_i \rangle$ and derive a lower bound of the incomplete log-likelihood with the parameter $\vartheta^{(t)}$ in the $t$-th iteration, formulated as follows:

$$\log(p(D_i | \vartheta)) = \log \sum_{k=1}^{K} \frac{\prod_{j=1}^{R} \gamma_i^j}{\prod_{j=1}^{R} \alpha_k^{\gamma_i^j}} \cdot \prod_{j=1}^{R} (1 - \alpha_k^{\gamma_i^j})^{1-\gamma_i^j}$$

We adopt Expectation-Maximization algorithm (term EM, including Expectation step and Maximization step) to find maximum likelihood estimates of parameters in Eq. (3). 

#### E-step. We maximize the log-likelihood by maximizing its lower bound. Before that, we compute the expectation for probability distribution of latent variables in the $t$-th iteration, denoted by $\mu_{ki}^{(t)}$ as follows:

$$\mu_{ki}^{(t)} = \frac{\prod_{j=1}^{R} \gamma_i^j}{\sum_k \prod_{j=1}^{R} \gamma_i^j} \cdot \prod_{j=1}^{R} (1 - \alpha_k^{\gamma_i^j})^{1-\gamma_i^j}$$

### TABLE I: Main notations and definitions.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D = {&lt;^v x_i, ^t x_i, ^v y_i&gt; \mid i = 1, \ldots, N}$</td>
<td>N samples on the visual, textual, and emoticon modality respectively.</td>
</tr>
<tr>
<td>$\langle y_i, z_i \rangle_k$</td>
<td>The $k$-th situation formed by the outputs of visual sentiment, textual sentiment and sentiment consistency.</td>
</tr>
<tr>
<td>$\gamma_i^j$</td>
<td>The occurrence (0 or 1 value) of the $j$-th emoticon in the $i$-th tweet.</td>
</tr>
<tr>
<td>$^v\theta$</td>
<td>Parameters of the visual CNN.</td>
</tr>
<tr>
<td>$^t\theta$</td>
<td>Parameters of the textual DCNN.</td>
</tr>
<tr>
<td>$\alpha_k$</td>
<td>The SC-weight of the $k$-th sentiment situation with the $j$-th emoticon label observed.</td>
</tr>
<tr>
<td>$\mu_{ki}^{(t)}$</td>
<td>The expectation for probability distribution of latent variables in the $t$-th iteration.</td>
</tr>
</tbody>
</table>
where $\theta = \{v\theta, t\theta\}$ and we compute $p(\langle y_i, t_i, z_i \rangle_k | x_i; \theta^{(t)})$ from the forward propagation of visual and textual prediction, since $y_i$, $t_i$, and $z_i$ are independent. Then, we write the expected complete log-likelihood function as:

$$Q(\vartheta; t^{(t)}) = \sum_{i=1}^{N} \mu_{ki}^{(t)} \log(p(\langle y_i, t_i, z_i \rangle_k | x_i; \vartheta)).$$  \hspace{1cm} (5)

M-step. To maximize the log-likelihood with the expected parameter $t^{(t)}$ in the $t$-th iteration, we write the derivative of $Q$ w.r.t. $\vartheta$ as follows:

$$\frac{\partial Q}{\partial \vartheta} = \sum_{k=1}^{K} \sum_{v=0}^{V} \mu_{ki}^{(t)} \frac{\partial}{\partial \vartheta} \log(p(\langle y_i, t_i, z_i \rangle_k | x_i; \vartheta)).$$  \hspace{1cm} (6)

According to the conditional probability equation, we split the joint probability distribution into two parts with $\alpha$ and $\vartheta$ as conditional variables:

$$p(\langle y_i, t_i, z_i \rangle_k | x_i; \theta, \alpha_k) = p(\langle y_i \rangle_{\alpha_k} | \langle y_i \rangle_{\alpha_k}, \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta) \cdot p(\langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k | x_i; \theta).$$  \hspace{1cm} (7)

The gradient of $Q$ w.r.t $\vartheta$ can be split into two parts: $Q$ w.r.t $v\vartheta$ and $Q$ w.r.t $t\vartheta$. We take $Q$ w.r.t $v\vartheta$ for example, which can be computed according to Eq.(6) and Eq.(7) as:

$$\frac{\partial Q}{\partial \vartheta} = \sum_{k=1}^{K} \mu_{ki}^{(t)} \frac{\partial}{\partial \vartheta} \log(p(\langle y_i, t_i, z_i \rangle_k | x_i; \vartheta))$$

$$= \sum_{v=0}^{V} \left[ \sum_{k=1}^{K} \mu_{ki}^{(t)} \frac{\partial}{\partial \vartheta} \log(p(\langle y_i \rangle_{\alpha_k} | \langle y_i \rangle_{\alpha_k}, \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta)) \right]$$

$$+ \left( \sum_{k=1}^{K} \mu_{ki}^{(t)} \frac{\partial}{\partial \vartheta} \log(p(\langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k | x_i; \theta)) \right).$$  \hspace{1cm} (8)

The equation above satisfies $\sum_{v=0}^{V} \left( \sum_{k=1}^{K} \mu_{ki}^{(t)} \frac{\partial}{\partial \vartheta} \log(p(\langle y_i \rangle_{\alpha_k} | \langle y_i \rangle_{\alpha_k}, \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta)) \right) = 1$ and $\sum_{k=1}^{K} \mu_{ki}^{(t)} = 1$, which can be regarded as the process of minimizing the cross-entropy between the estimated ground truth distribution and the predicted distribution from forward propagation. Especially, since we utilize Eq.(12) and Eq.(13) to compute $p(\langle z_i \rangle_k = 0 | x_i; \theta)$ in the forward propagation, we get the gradient of consistency w.r.t $v\vartheta$ as:

$$\frac{\partial}{\partial \vartheta} \log(p(\langle z_i \rangle_k = 0 | x_i; \vartheta)) = (1 - \sigma) \frac{\partial p(\langle y_i \rangle_{\alpha_k} | \langle y_i \rangle_{\alpha_k}, \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta)}{\partial \vartheta}$$

$$\times \left[ \frac{p(\langle y_i \rangle_{\alpha_k} | \langle y_i \rangle_{\alpha_k}, \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta^{(t)})}{p(\langle y_i \rangle_{\alpha_k} | \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta^{(t)})} + 1 - \frac{p(\langle y_i \rangle_{\alpha_k} | \langle y_i \rangle_{\alpha_k}, \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta^{(t)})}{p(\langle y_i \rangle_{\alpha_k} | \langle t_i \rangle_{\alpha_k}, \langle z_i \rangle_k; \theta^{(t)})} \right].$$  \hspace{1cm} (9)

where $y^v$ denotes the consistency variable output from the corresponding prediction component (see Section IV for details).

In the same way, we can get the gradient of $Q$ w.r.t $t\vartheta$ and the gradient of consistency w.r.t $t\vartheta$ for textual modality.

Additionally, utilizing Eq.(1), Eq.(5) and Eq.(7), we can compute the gradient of $Q$ w.r.t $\alpha$ as:

**Algorithm 1 Weakly Supervised Multi-modality Deep Learning on Model Training and Sentiment Prediction**

**Input**: The microblog tweets $D = \{\langle x_1, y_1, z_1 \rangle, \ldots, \langle x_N, y_N, z_N \rangle\}$ whose sentiment needs to be predicted.

**Output**: The polarity for microblog tweets sentiment prediction.

**Model Training**: 

**Step 1.** The CNN and DCNN models are pretrained by using “clean” labels. The initialized parameters of the multi-modal noise model are randomized generated.

**Step 2.** For each tweet, the sentiment combination $\langle y_i, t_i, z_i \rangle$ is predicted from the forward propagation of corresponding deep learning components;

**Step 3.** Implementing EM algorithm and Back Propagation algorithm to optimize the parameters of CNN as Eq.(8), DCNN like Eq.(8), and SC-weights as Eq.(11) on emotion-labels;

**Step 4.** Steps 2 and 3 above are repeat in each iteration until the convergence of loss.

**Sentiment Prediction**:

**Step 5.** Each microblog tweet is taken as the input of the retrained CNN and DCNN models for its visual and textual feature representations. The final prediction is produced by softmax regression over these modalities.

**Output**: The polarity for microblog tweets sentiment prediction.

**Details**:

We introduce the implementation details in this section. For visual sentiment prediction, AlexNet [21] based CNN model is adopted, which computes visual features by forward propagating a $227 \times 227$ RGB image through 5 convolutional layers and 2 fully-connected layers. We add a 3-output fully-connected layer for the visual sentiment and a 7-output fully-connected layer for the sentiment consistency. For textual sentiment prediction, the word embedding scheme is leveraged.

**Algorithm 1 Weakly Supervised Multi-modality Deep Learning on Model Training and Sentiment Prediction**

**Input**: The microblog tweets $D = \{\langle x_1, y_1, z_1 \rangle, \ldots, \langle x_N, y_N, z_N \rangle\}$ whose sentiment needs to be predicted.

**Output**: The polarity for microblog tweets sentiment prediction.

**Model Training**: 

**Step 1.** The CNN and DCNN models are pretrained by using “clean” labels. The initialized parameters of the multi-modal noise model are randomized generated;

**Step 2.** For each tweet, the sentiment combination $\langle y_i, t_i, z_i \rangle$ is predicted from the forward propagation of corresponding deep learning components;

**Step 3.** Implementing EM algorithm and Back Propagation algorithm to optimize the parameters of CNN as Eq.(8), DCNN like Eq.(8), and SC-weights as Eq.(11) on emotion-labels;

**Step 4.** Steps 2 and 3 above are repeat in each iteration until the convergence of loss.

**Sentiment Prediction**:

**Step 5.** Each microblog tweet is taken as the input of the retrained CNN and DCNN models for its visual and textual feature representations. The final prediction is produced by softmax regression over these modalities.

**Algorithm 1 Weakly Supervised Multi-modality Deep Learning on Model Training and Sentiment Prediction**

**Input**: The microblog tweets $D = \{\langle x_1, y_1, z_1 \rangle, \ldots, \langle x_N, y_N, z_N \rangle\}$ whose sentiment needs to be predicted.

**Output**: The polarity for microblog tweets sentiment prediction.

**Model Training**: 

**Step 1.** The CNN and DCNN models are pretrained by using “clean” labels. The initialized parameters of the multi-modal noise model are randomized generated;

**Step 2.** For each tweet, the sentiment combination $\langle y_i, t_i, z_i \rangle$ is predicted from the forward propagation of corresponding deep learning components;

**Step 3.** Implementing EM algorithm and Back Propagation algorithm to optimize the parameters of CNN as Eq.(8), DCNN like Eq.(8), and SC-weights as Eq.(11) on emotion-labels;

**Step 4.** Steps 2 and 3 above are repeat in each iteration until the convergence of loss.

**Sentiment Prediction**:

**Step 5.** Each microblog tweet is taken as the input of the retrained CNN and DCNN models for its visual and textual feature representations. The final prediction is produced by softmax regression over these modalities.
by using Word2Vec tool [45]. Subsequently, the DCNN [23] is adopted to learn a high-level representation, which consists of 3 convolutional layers with a special dynamic k-max pooling layer, together with a fully-connected layer. We add a 3-output fully-connected layer for the textual sentiment and a 7-output fully-connected layer for the sentiment consistency. It’s noted that we omit the analysis of word order in word embedding. The reasons is two-fold: On the one hand, we use the advanced off-the-shelf word2vec tool. On the other hand, the width of filters in the first convolution layer of DCNN is set to be large enough, which can learn to recognize specific n-grams as stated in [23]. Note that, any other cutting-edge techniques in textual feature representation can be also leveraged here, while our contribution is orthogonal to this step. Indeed, our work may mainly focus on the part of weakly supervised modeling, while both DCNN and CNN are directly used as they are off-the-shelf schemes in the corresponding textual and visual channels. Before training the model with noisy labels, we first pre-train a “clean” model in a small scale setting by using a limited number of clean labels (in our implementation we have over 6K clean labels and over 80K noisy labels).

As shown in Fig.2, we compute the consistency of the predicted sentiments from two modalities respectively by a softmax function. Using KL divergence, we define the evaluation function $D$ for the two output distributions, $p$ and $q$ (from two prediction models) as below:

$$D(p||q) = D_{KL}(p||q) + D_{KL}(q||p).$$

(12)

The final consistency probability distribution $p(z_i = 0|x_i; \theta)$ can be computed by a Sigmoid function as:

$$p(z_i = 0|x_i; \theta) = \sigma(D(p(y^{v}_i|x_i; ^v\theta)||p(y^{v}_i|x_i; ^t\theta))),$$

(13)

where, $y^{v}_i$ and $y^{v}_i$ are output variables of the visual and textual prediction components, respectively. We aim to describe more diverse sentiment situations to refine the influence range of each modality to the overall label, a.k.a., the emoticon label in the scenario of our application.

V. EXPERIMENTS

A. Dataset and Preprocessing

Dataset. We constructed a large-scale dataset for multi-modal microblog sentiment prediction, which is crawled from Sina Weibo. This dataset is collected according to the 10 hottest topics from Sep. 2013 to Feb. 2014, covering trivial affairs, weather, work, stars, films, and international events. There are totally more than 1.4 million microblogs in the dataset. We remove duplicate microblogs and obtain 435,458 tweets. The percentage of tweets on textual modality, visual modality, emoticon modality and the combination of the three modalities is shown in Tab.III. At the end, we retain 80K tweets with textual, visual and emoticon contents, as our noisy “weak” labels for training. To collect clean labels, we first manually label the sentiment polarities for over 6K tweets by three expert annotators. Then we conduct two-round consistency evaluations of annotations by these experts to evaluate the quality of over 6K clean labels as shown in Tab.V. Between the first and the second rounds, these experts exchange views and unify opinions to reduce subjective errors. At the end of the second round, “N. of T.” is reduced to 0 by the majority voting rule. The principle of the above annotation is based on a well-known emotion model, i.e., Plutchik’s Wheel of Emotions, which is widely used in the psychological studies [46]. Following the same way, we manually label the sentiment polarities of both visual and textual channels for these tweets. Finally, we obtain over 6K tweets with clean labels (4,196 positive tweets, 1,354 negative tweets and 621 neutral tweets), where 5,271, 400, and 500 tweets are selected as the training set, the validation set and the testing set, respectively. Some exemplar tweets with sentiment labels are shown in Fig.5.

Preprocessing. During the preprocessing period, we first remove duplicate microblogs and obtain 435,458 tweets. Then, we filter the hashtags and external links. Finally, we segment sentence into words by using Chinese auto-segmentation system ICTCLAS [47]. For the sake of simplicity, we also replace the unmatched words with the sign “unknown” according to the commonly-used Chinese dictionary.

B. Baselines

We compared the proposed WS-MDL with three baselines for multi-modal sentiment prediction. It’s noted that WS-MDL is trained based on visual, textual and emoticon modalities jointly. To conduct a fair comparison, we compare the recent methods that are also based on the above three modalities. 1) CBM-LR [24]: A cross-media bag-of-words model based on Logistic Regression (termed CBM-LR) for multi-modal sentiment prediction. In CBM-LR, the texts and emoticons are modeled as bag-of-words, and the high-level semantic features of images are extracted by Adjective Noun Pair (ANP) detectors [26], which are treated as visual bag-of-words. These multi-modal bag-of-words are then concatenated as the input of Logistic Regression model for training. 2) CBM-SVM [24]: A cross-media bag-of-words model based on linear Support Vector Machine (termed CBM-SVM) for
multi-modal sentiment prediction. In CBM-SVM, the texts, images and emoticons are modeled to compose multi-modal bag-of-words in the same way as CBM-LR. Linear Support Vector Machine are trained on these bag-of-word features for multi-modal sentiment prediction. 3) HGL [17]: A multi-modal hypergraph learning method (termed HGL) for multi-modal sentiment prediction. In this method, the features of each modality are extracted in the same way as CBM-LR. HGL first computes the pairwise tweet similarities on different modalities respectively, and then construct a multi-modal hypergraph model, where each vertex represents a tweet and the hyperedges are formed by the “centroid” vertex and its k-nearest neighbors on different modalities. Alternating optimization is leveraged to learn the final relevance score of each tweet. So far, HGL has achieved the state-of-the-art performance for multi-modal microblog sentiment prediction.

C. Comparisons on Sentiment Prediction

We compare WS-MDL to above state-of-the-art methods, all of which are based on supervised learning, including CBM-LR, CBM-SVM and HGL. Note that our method treats the emoticon modality as noisy labels, and the final prediction is produced by softmax regression over three modalities. In contrast, the aforementioned baselines all treat the emoticon modality as one of the feature channels. Tab.VI shows the performance comparison between WS-MDL and baselines on multi-modal sentiment prediction, from which we can observe that WD-MDL outperforms the comparison baselines in accuracy. Note that we use extra noisy data for training. However, all baselines cannot use such noisy data in the same way, and instead, “clean” labels are the required input for these supervised methods.

To evaluate the computation efficiency of the proposed model, we compare WS-MDL to above state-of-the-art methods by weakly-supervised training way on the same scale of data. We first roughly divide the emoticons into three sentiment polarities (positive, negative and neutral) as the noisy labels. Then we retrain CBM-LR and CBM-SVM on the noisy labels. Note that we leave HGL out due to large memory and computational consumptions. Tab.VII shows the performance comparison between WS-MDL and baselines on weakly-supervised multi-modal sentiment prediction, from which we can observe that WD-MDL outperforms the comparison baselines in accuracy. Besides, the comparison results from Tab.VI and Tab.VII shows that, by using the large noisy data, CBM-LR and CBM-SVM both gets worse, say from 3.9% to 4.3% accuracy decrease comparing to that by using clean data, while after adding noisy data into the training of WS-MDL, both visual and textual components gets better as shown in Fig. 7 (b). It reflects that the comparison methods except for WS-MDL hardly benefit from the large noisy data.
Fig. 6: Performance comparisons on textual, visual modality, and overall modalities for CBM-LR, CBM-SVM, HGL and WS-MDL method.

We further compare the sentiment prediction performance within individual modality, respectively in Fig.6. Our method also achieves the best performance with respect to all modalities. It is noted that WS-MDL is less superior than HGL on visual modality. It can be explained by the fact that HGL is based on ANP detectors [26], which is widely regarded as the best performed visual sentiment predictors. ANP detectors are not suitable for WS-MDL due to their fixed models, which is unable to be retrained for the integral multi-modal scheme. Indeed, our WS-MDL is open to all other cutting-edge detectors (positive, negative and neutral) are obtained in a reasonable extent, which further explains the effectiveness of WS-MDL. We also compare the performances of visual model (CNN) and textual model (DCNN) on different training periods, shown in Fig.7 (b). As we can see, models trained along with the proposed multi-modal noise inference scheme outperform the others, which reflects its contribution to the optimization of visual and textual models. Moreover, we can also find that the overall training process contributes more to the optimization of visual model, which obtains 17% promotion compared to 3% of the textual one. It can be explained by the fact that the initial visual model is based on the pretrained AlexNet model [21] while the parameters of the initial textual model is randomly initialized (There is no pretrained model for DCNN on the microblog dataset currently).

In addition, we provide the loss curve of training and the accuracy curve of validation in Fig. 8 to trace the internal performance at each epoch. Fig. 8 (a) verifies the convergence of the proposed model and Fig. 8 (b) shows that WS-MDL reaches the optimum iteration at the 37-th epoch. Furthermore, in Fig. 8 (b), we can find that the visual component benefits more from the model than the textual component. It’s probably due to sparse valuable information and more noises in the textual channel. We can also find that overall performance slightly benefits from the promotions of the textual component and the visual component, which reflects the relevance between the model performance and the component performances.

Fig. 7: Detailed Comparisons inside WS-MDL. (a) Comparison on precision and recall of different sentiment categories. (b) Comparison on different training periods (T-Modality and V-Modality mean textual and visual modalities, respectively).

D. Quantitative Analysis of WS-MDL

We further compare the sentiment prediction performance of WS-MDL. In particular, we first compare the precision and recall for different sentiment categories, as shown in Fig.7 (a). The precision and recall of three sentiment categories (positive, negative and neutral) are obtained in a reasonable extent, which further explains the effectiveness of WS-MDL. We also compare the performances of visual model (CNN) and textual model (DCNN) on different training periods, shown in Fig.7 (b). As we can see, models trained along with the proposed multi-modal noise inference scheme outperform the others, which reflects its contribution to the optimization of visual and textual models. Moreover, we can also find that the overall training process contributes more to the optimization of visual model, which obtains 17% promotion compared to 3% of the textual one. It can be explained by the fact that the initial visual model is based on the pretrained AlexNet model [21] while the parameters of the initial textual model is randomly initialized (There is no pretrained model for DCNN on the microblog dataset currently).

In addition, we provide the loss curve of training and the accuracy curve of validation in Fig. 8 to trace the internal performance at each epoch. Fig. 8 (a) verifies the convergence of the proposed model and Fig. 8 (b) shows that WS-MDL reaches the optimum iteration at the 37-th epoch. Furthermore, in Fig. 8 (b), we can find that the visual component benefits more from the model than the textual component. It’s probably due to sparse valuable information and more noises in the textual channel. We can also find that overall performance slightly benefits from the promotions of the textual component and the visual component, which reflects the relevance between the model performance and the component performances.

E. Visualization of Sentiment Predictions

We further investigate two details in multi-modal sentiment prediction, a.k.a., whether the sentiment can be conveyed in some spatial locations in image or text, and whether the multi-modal information is truly identical. To answer the above two questions, we visualize the feature maps of the 5-th layer of CNN (the last convolutional layer for semantic representation) and the 1-st layer of DCNN. And the reason we choose the first convolutional layer of DCNN is that it can better present the origin structure of the sentence inputted without the effect of dynamic k-max pooling [23].

The visualization results of feature maps are shown in Fig.9, where we present the feature maps with the strongest activation as inspired by [48]. To be specific, we draw the heatmap by the activations in each feature map for CNN, while the words with the strongest activation in the textual feature maps are marked with red highlights. The areas covered by strong activations of the feature maps well reflect the sentiment information, e.g., the blood on the face (the second example) could be the factor of negative sentiment. It indicates that our optimized deep CNN can effectively learn the factor of visual sentiment by the spatial locations of strongest sentiment responses.
Every time I listen to the songs of Wenlan, I am distracted. It is really for smart kids to listen.

Yahoo, send a shooting slinkingly! These days how about taking some visual pictures and sidelights?

Boy, you have grown up, you should be sensible. Yeah, so now you are the “big” man.

Another sleepless night to see the TV show LaMaZhenZhuan. How do I handle this? Will I be weak?

LaMaZhenZhuan, a positive energy, I like Muzi and Xiabin.

Fig. 9: The focus visualization of sentiment prediction in our method. (a) The input images, (b) the strongest feature map in layer 5 of CNN, (c) the strongest feature map in layer 1 of DCNN, (d) the emoticons and (e) the output sentiments.

Additionally, Fig. 9 presents the matrix of the feature map with strongest activation in DCNN. According to the activations of textual feature maps, the corresponding words with strong activations are marked with red highlights. It is easy to find that the words with red highlights carry specific sentiment factors, which are mostly indispensable component for textual sentiment presentations. All these reflect the effectiveness of our optimized DCNN on inferring the potential textual sentiment factors. Fig. 9 further reflects the inner relevance of the multi-modal consistency on sentiment representation. Taking the first line for example, there is almost not semantic correlation between the textual and the visual content, but our WS-MDL model can automatically learn the most explainable relevance, i.e., the visual object cat and the textual entity kid. However, we can also find that the emoticons of the 1-st and the 3-rd instances are not so consistent with the final sentiment results and the visual and textual focuses (in column b and column c respectively) are relatively not on the best positions of images and texts, respectively. It might be caused due to the fact that the images or texts of these two instances have relatively strong sentiments which are hardly affected by

4The Chinese sentences of instances have been translated into English for understanding.

Fig. 11: The correlation between top 100 emoticons with high frequency and 9 sentiment combinations according to SC-weight.

emoticons during training.

F. On SC-weights for Inferring Label Noise

In our scheme, the emoticon labels serve as the noisy labels to be used and evaluated during training, whose confidence is iteratively updated its sentiment contribution weight (SC-weight) during WS-MDL training. The distribution for different sentiments reflects whether the model is trained well.
Fig. 10: The SC-weight of selected top 100 emoticon-labels when textual sentiment and visual sentiment are both positive (termed TV positive), negative (termed TV negative), and neutral (termed TV neutral), respectively.

TABLE VIII: Performance comparison of different emoticon selecting manners

<table>
<thead>
<tr>
<th>Emoticon selection</th>
<th>Num.</th>
<th>Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random-selected#1</td>
<td>100</td>
<td>0.662</td>
</tr>
<tr>
<td>Random-selected#2</td>
<td>100</td>
<td>0.666</td>
</tr>
<tr>
<td>Random-selected#3</td>
<td>100</td>
<td>0.676</td>
</tr>
<tr>
<td>PCA</td>
<td>100</td>
<td>0.709</td>
</tr>
<tr>
<td>LDA</td>
<td>2</td>
<td>0.751</td>
</tr>
<tr>
<td>Weight-based</td>
<td>100</td>
<td>0.734</td>
</tr>
</tbody>
</table>

In this paper, we study the challenging problem of predicting multi-modal sentiments for social media tweets. Our main contribution is to train a discriminative model from cheaply available emoticon labels for multi-modal prediction. This is achieved by a novel weakly-supervised multi-modal deep learning (WS-DML) framework. In particular, we firstly compute the sentiment probability distributions and the multi-modal sentiment consistency from the pretrained CNN and DCNN models. Then, we train a probabilistic graphical model to distinguish the contribution weights of the noisy label, which are further sent back to update the parameters of CNN and DCNN models, respectively. Experimental comparisons to the state-of-the-art methods demonstrate that our method has achieved the state-of-the-art performance on the multi-modal sentiment prediction task. Along with this work, a dataset for multi-modal sentiment prediction is further released, which is the largest one in the literatures.

Many potential improvements are there to make a better usage of noisy labels. For example, the noisy label could be user’s online behavior or logs if being accessible from the social media website. Besides, in our future work, we will further investigate the order of the emoticon labels, which can be added as a constraint in the WS-MDL model.

ACKNOWLEDGMENT

This work is supported by the National Key R&D Program (No. 2016YFB1001503), the Special Fund for Earthquake Research in the Public Interest No.201508025, the Nature Science Foundation of China ( No. 61422210, No. 61373076, No. 61402388, and No. 61572410), and the Xiamen Science and Technology Project(No.3502ZZ20153003).

REFERENCES


Fuhai Chen received the B.S. degree from the Department of Cognitive Science in Xiamen University, China, in 2014. He is now a Ph.D. candidate pursuing master-doctor degree in the Department of Cognitive Science at Xiamen University now. His research interests include machine learning, multimedia and computer vision.

Rongrong Ji(SM14) is currently a Professor with Xiamen University, Xiamen, China, where he directs the Intelligent Multimedia Technology Laboratory and serves as a Dean Assistant with the School of Information Science and Technology. Before that, he was a Post-Doctoral Research Fellow with the Department of Electrical Engineering, Columbia University, New York, NY, USA, from 2010 to 2013. He received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China. He was a Visiting Student with the University of Texas at San Antonio, San Antonio, TX, USA, a Research Assistant with Peking University, Beijing, China, and a Research Intern with Microsoft Research Asia, Beijing. He is the Guest Editor of the IEEE MULTIMEDIA MAGAZINE, Neurocomputing, Signal Processing, ACM Multimedia Systems, and Multimedia Tools and Applications, a Special Session Chair of the 2014 Indian Council of Medical Research, the 2013 IEEE Visual Communications and Image
Processing Conference, the 2013 Magnetism and Magnetic Materials Conference, and the 2012 Pacific-Rim Conference on Multimedia, a Program Committee Member of more than 30 flagship international conferences, including the IEEE Conference on Computer Vision and Pattern Recognition in 2013, the International Conference on Computer Vision in 2013, and the ACM Multimedia from 2014 to 2010, and a reviewer for over 20 IEEE/ACM Transactions such as TPAMI and TKDE.

Jinsong Su received the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2011. He is now an associate professor in the Software School of Xiamen University. His research interests include natural language processing and social computing.

Donglin Cao is currently an assistant professor at the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, China. His research interest is cross-media information retrieval.

Yue Gao (SM14) received the B.E. degree in Department of Electronic Information Engineering, Harbin Institute of Technology, the Master Degree and the Ph.D. degree in School of Software and Department of Automation of Tsinghua University, respectively. He is currently an associate professor in School of Software and TNList, Tsinghua University, Beijing, China. He has been working in School of Computing, National University of Singapore and School of Medicine, University of North Carolina at Chapel Hill from 2012 to 2016. He is the recipient of the 1000 Youth Talent Plan Grant of China. He has published 80+ papers in premier journals and conferences like TIP, TMM, TCSVT, TNNLS, TOMM, TGRS, TIE, Human Brain Mapping, MICCAI, CVPR, IJCAI, and ACM Multimedia. His research falls in the field of computer vision, machine learning, medical image analysis, and software engineering. His recent research work contains 3D understanding (retrieval/recognition), brain image analysis, and software defect prediction.