When E-commerce Meets Social Media: Identifying Business on WeChat Moment Using Bilateral-Attention LSTM

Tianlang Chen
University of Rochester
tchen45@cs.rochester.edu

Yuxiao Chen
University of Rochester
ychen211@cs.rochester.edu

Han Guo
Institute of Computing Technology, Chinese Academy of Sciences
guohan@ict.ac.cn

Jiebo Luo
University of Rochester
jluo@cs.rochester.edu

ABSTRACT

WeChat Business, developed on WeChat, the most extensively used instant messaging platform in China, is a new business model that bursts into people's lives in the e-commerce era. As one of the most typical WeChat Business behaviors, WeChat users can advertise products, advocate companies and share customer feedback to their WeChat friends by posting a WeChat Moment–a public status that contains images and a text. Given its popularity and significance, in this paper, we propose a novel Bilateral-Attention LSTM network (BiATT-LSTM) to identify WeChat Business Moments based on their texts and images. In particular, different from previous schemes that equally consider visual and textual modalities for a joint visual-textual classification task, we start our work with a text classification task based on an LSTM network, then we incorporate a bilateral-attention mechanism that can automatically learn two kinds of explicit attention weights for each word, namely 1) a global weight that is insensitive to the images in the same Moment with the word, and 2) a local weight that is sensitive to the images in the same Moment. In this process, we utilize visual information as a guidance to figure out the local weight of a word in a specific Moment. Two-level experiments demonstrate the effectiveness of our framework. It outperforms other schemes that jointly model visual and textual modalities. We also visualize the bilateral-attention mechanism to illustrate how this mechanism helps joint visual-textual classification.

KEYWORDS

attention model, joint visual-textual learning, multimodality analysis, WeChat business

ACM Reference Format:
Tianlang Chen, Yuxiao Chen, Han Guo, and Jiebo Luo. 2018. When E-commerce Meets Social Media: Identifying Business on WeChat Moment Using Bilateral-Attention LSTM. In *The 2018 Web Conference Companion, April 23–27, 2018, Lyons, France,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3184558.3186346

WWW' 18 Companion, April 23–27, 2018, Lyons, France

ACM ISBN 978-1-4503-5640-4/18/04.
https://doi.org/10.1145/3184558.3186346

1 INTRODUCTION

Figure 1: Examples of WeChat Moment. A red dotted box indicates a WeChat Business Moment, a blue dotted box indicates a WeChat non-business Moment.

In the e-commerce era, the rise of WeChat Business can be regarded as an significant event in the Chinese e-commerce history. WeChat Business, developed on WeChat, one of the most well-known messaging platforms with 806 millions monthly active users[1] in China, is a new business mode that refers to seller's advertising and trading activities on WeChat. Typically, as one of the most popular and effective promotion strategy, sellers can advertise their products, companies platforms and other services via WeChat Moments.

Like Instagram, WeChat Moment is a platform where users can post text and pictures, which can be accessed and commented by their WeChat friends. In a Moment, sellers can effectively advertise their products or services, share the customer feedbacks and show their trading achievements. A Moment related to these topics should be regarded as a WeChat Business Moment. Figure 1 shows several examples that belong to WeChat Business Moments. Our goal is to create a high-performance WeChat Business Moment classifier that can accurately identify WeChat Business Moments, based on their image and text contents. For online shopping addicts, they need the push notification service from these Moments to provide instant and comprehensive information for potential purchases. However, for online shopping haters, they want to block these Moments because they are nuisance. A robust classifier is beneficial for both kinds of users. It can identify the WeChat Business Moments from the Moment pool and facilitate appropriate actions by different kinds of users.

[1] http://tech.qq.com/a/20160518/067853.htm

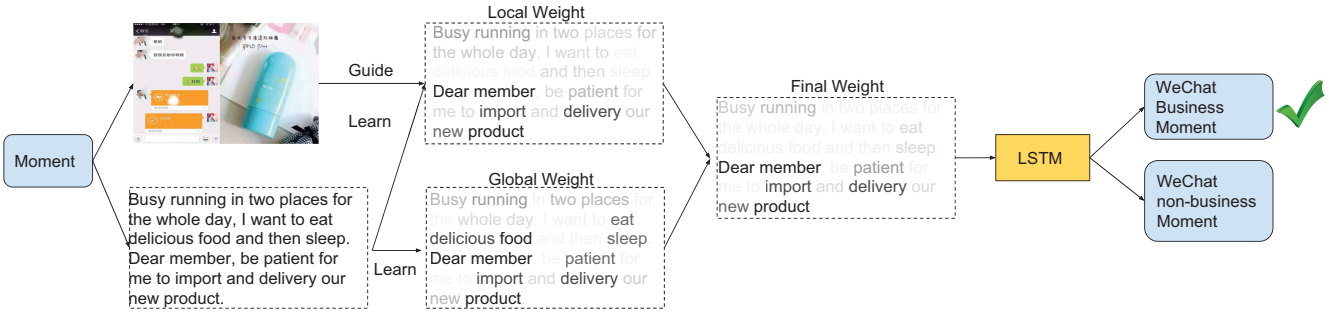

Figure 2: Overview of the proposed work.

We formally define our task as a visual-textual binary classification task to classify WeChat Business Moments and WeChat non-business Moments, where each Moment may contain a text message and multiple images (up to 9). To accurately identify WeChat Business Moments, we propose a novel bilateral-attention LSTM network. Different from previous works that equally consider visual and textual features and build fusion-based models, our model is based on text classification using LSTM and we make the best use of weak image information by creating an image-guide attention mechanism. In particular, we believe that for a word in a specific Moment, there are two significant weights to measure its importance in its associated sentence, namely the global weight and local weight. The global weight reflects the overall importance of a word for the classification task and is insensitive to the corresponding images of the Moment the word belongs to. On the other hand, the local weight reflects the local importance of a word in a specific Moment that is related to the word's corresponding images. In other words, same words in different Moments possess same global weights but different local weights. The final weight of a word should be the combination of its global weight and local weight. Figure 2 shows the framework of our model. When we predict whether a Moment is related to WeChat Business, a self-learned attention mechanism will learn the global weight of each word, and an image-guide attention mechanism will further figure out the local weight of each word from the Moment's specific image environment. In Figure 2, "eat", "delicious", "food", "sleep", "Dear member", "patient", "import", "delivery" and "product" have high global weights as they are significant words for the WeChat Business Moment classification task. However, combined with the local weight of each word, only "Dear member", "import", "delivery" and "product" hold the highest final weights since the images are related to WeChat screenshot and cosmetics. In the end, the Moment can be correctly predicted as a WeChat Business Moment by the LSTM network.

We make the following contributions in this work:

- We propose an end-to-end bilateral-attention LSTM model that can successfully capture the global and local importance of a word in a specific Moment. To figure out the local weight by an image-guide attention mechanism, we propose an efficient method to accurately classify WeChat Moment images into categories in a semi-supervised fashion.
- We perform two-level experiments on a WeChat Moment dataset to demonstrate the effectiveness of our framework. In particular, we demonstrate that the image-guide attention mechanism makes better use of image information compared with other joint visual-textual learning models for our task. We also visualize the bilateral-attention mechanism on significant examples to illustrate how it works.

2 RELATED WORK

Recently, in keeping up with the wave of e-commerce, many researchers have paid attention to social network business and advertising. For instance, by analyzing more than 7000 tweets regarding the Fortune 500 companies, Swani et al. conclude that different branding and selling strategies exist in B2B and B2C settings, such as in terms of message appeal, cues, links, and hashtags [14]. Zhai et al. build an RNN network, which maps queries and ads to real valued vectors so that one can easily compute the quality of a (query, ad) pair [23].

Our work also tracks the popular e-commerce research while integrating a new attention model. The attention model has been applied to different research topics and tasks in recent years, with its strong capacity to capture key words or local image regions that can provide more significant information. For image captioning, Xu et al. [17] first propose an attention-based model that automatically learns to describe the content of images. You et al. [20] propose a novel framework that combines the top-down and bottom-up approaches through a semantic attention model. In video captioning, Guo et al. [5] propose a novel end-to-end attention-based LSTM framework with semantic consistency, to transfer videos to natural sentences. In addition, in the field of image question answering, Yang et al. [18] present stacked attention networks that could learn to answer natural language questions from images. Shih et al. [13] create an attention-based model that learns to answer visual questions by selecting image regions relevant to the text-based query. Inspired by these results, we propose a novel model that combines two kinds of attention mechanism with strong interpretability for our joint visual-textual classification task. We demonstrate that compared with other models, the weak image feature could exert its influence in a better way as a guidance to adjust the text word weight.

Although most of the recent social media data mining research mainly focuses on western social media services, such as Twitter, Facebook and Instagram, researchers start to pay attention to WeChat due to its high popularity in China. For instance, Wang et al. investigate how WeChat usage reinforces, reconfigures, and enhances existing Chinese social practices. They propose a new theoretical concept, space collapse [15]. Zang et al. analyze the growth patterns of Wechat online social network and propose a NetTide Model to fit the growth [22]. Qiu et al. analyze the growth, evolution and diffusion patterns of WeChat Messaging Group [11], and Li et al. discover the diffusion patterns of information in Moments by tracking a large amount of pictures in Moment [9].

3 BIATT-LSTM

3.1 Basic LSTM for text classification

For our task, a basic LSTM network can be implemented to identify a WeChat Business Moment by classifying its associated text, it receives the text of each Moment as input sample and predict its class. This basic LSTM network could be represented as the blue part in Figure 3. It contains a word-embedding layer that maps each word into a feature vector and a LSTM layer that extracts the hidden state of each time after inputting a new word, we could extract the last time's hidden state (h_T) or all times' hidden states (h_1,h_2,...,h_T) to represent the high-level feature of the input sentence. In the end, it transforms this high-level feature via several fully-connected layers and a softmax layer with 2 nodes that output the predicted probabilities of WeChat Business Moment and WeChat non-business Moment.

3.2 Self-learned attention VS Image-guide attention

As demonstrated in [19] [18], each word should possess a unique weight in a sentence, for a classification or regression task based on text, incorporating the attention mechanism to capture appropriate importance of each word will improve the model's performance. For our task, we consider that a word's attention weight in a specific Moment should be a two-level concept, in particular, the final attention weight of a word in a Moment should be the combination of a global weight and a local weight. The global weight of a word represents the overall importance and capacity of the word to classify samples in the classification task, it remains unchanged in different Moments. For example, both words "customer" and "mountain" should owe high global weights, because both of them are significant words to classify a text. The word "customer" has strong positive correlation with WeChat Business, a Moment whose text includes this word is usually a WeChat business Moment. In contrast, the word "mountain" has strong positive correlation with WeChat non-business. It should be noticed that this global weight is insensitive to Moments, in other words, the word "mountain" has the same global weight in different texts of different Moments. On the other hand, each word "mountain" in a specific Moment has a unique local weight that represents its importance in this Moment, based on the image environment of the Moment. For example, when "mountain" exists in a text whose corresponding images are completely related to ads, living goods, posters or foods, it should have a low local weight since it should not be the keyword the Moment intends to express. In contrast, if its corresponding images are mountain photos, landscape photos or even outdoor selfies, it should have a high local weight in the text. In the end, the final weight of a word in a text should be the combination of its global weight and local weight.

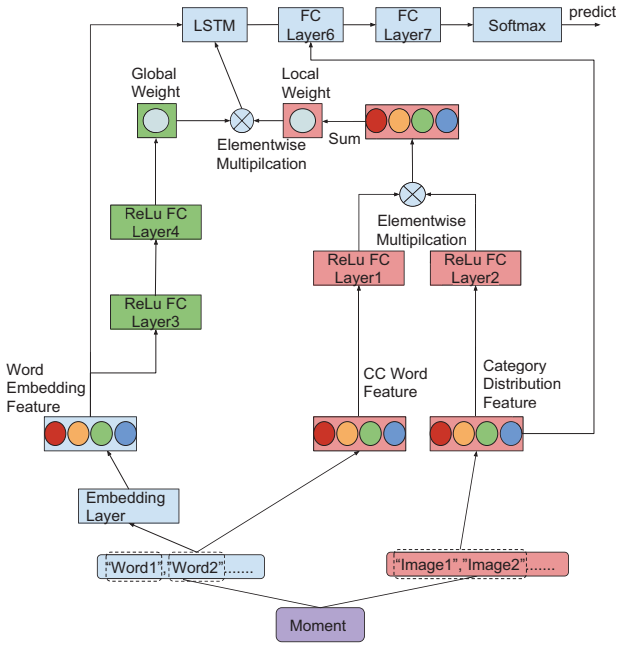

Figure 3: The structure of the proposed BiATT-LSTM. It contains a basic LSTM network (blue), a self-learned attention sub-network (green) and an image-guide attention sub-network (red).

For a given word, to figure out its local weight in a specific Moment, we construct two kinds of feature, one reflects the image environment of the corresponding Moment and one represents the word's relation with specific image environments. In particular, first, for all images in our dataset, we implement a semi-supervised classification framework to accurately classify images into n categories as a basis of two kinds of features. The whole process will be described in Section 3.3. Then, for a word W and a category C, we construct a correlation coefficient between them based on a Bayesian model and denote it as Θ_{WC}. In particular, if a word W belongs to a Moment, we define it as an occurrence of word W, and if a word W belongs to a Moment which contains images whose category is C, we define it as an occurrence of a word-category pair (W, C) . If a word occurs more than one time in a text, we only record it for one time. After we go through all texts in the train set, we can denote the posterior probability of a category C when we observe a word W by:

$$\Pr(C|W) = \frac{O(C,W)}{O(W)} \tag{1}$$

where $O(W)$, $O(C,W)$ represent the total occurrence number of a word W and a word-category pair (W, C). We also record the prior probability of the category C from the train set and denote it as $Pr(C)$. We could thus figure out the correlation coefficient Θ_{WC}

between a category C and a word W by:

$$\Theta_{WC} = \frac{Pr(C|W) - Pr(C)}{Pr(C)} \tag{2}$$

For each word, we construct a n-dimensional feature vector that records the correlation coefficient between the word and each specific image category and we call it category correlation word feature (CC word feature) and denote it as Θ_W. For a word in test set, if it occurs in the train set, we directly get its CC word feature. Otherwise, we use word2vec [10] to predict its five most similar words in the train set, and represent its feature by the mean value of these five words' CC word features. Meanwhile, for each Moment, focusing on its contained images, we construct a n-dimensional binary category distribution feature vector denoted as D_C. If there exists at least one image that belongs to a specific category, we set the corresponding value of the feature vector for this category as 1, otherwise, it is set to 0. In the end, for an input Moment, the local weight g_l of a word W is computed as:

$$g_l = sigmoid(\sum((W_{WC}\Theta_W) \odot (W_{DC}D_C))) \tag{3}$$

where Θ_W and D_C are the word's CC word feature and its corresponding images' category distribution feature in the same Moment. W_{WC} and W_{DC} are two matrices that map Θ_W and D_C to appropriate feature space. They are synchronously learned with the whole network. In our experiments, we set their dimensions as n x 200 which leads to the best performance. In the end, we compute the inner product of two vectors with an sigmoid transformation as the local weight of the word, sigmoid transformation restrict the weight's range from 0 to 1. In the whole process, the category distribution feature plays role as a filter, it strengthens the words express similar semantic content with the images and weaken the words express dissimilar semantic. Since this local weight is guided by the image information, we call the mechanism as image-guide attention.

For the global weight of a word, we train a two layers subnetwork with sigmoid unit at the top, it receives the embedding feature vector of a word as input and outputs a value in the range of $(0, 1)$ that represents the word's global weight. In particular, the global weight g_g of a word W is computed as:

$$\begin{aligned} h_1 &= ReLU(W_1 X_W) \\ g_g &= sigmoid(W_2 h_1) \end{aligned} \tag{4}$$

where X_W is the embedding feature vector of word W, W_1 and W_2 are learning matrices updated with the whole network. In our experiments, we set the dimensions of W_1 and W_2 as m x 200 and 200 x 1 to produce the best performance, where m is the dimension of word feature vector.

Note that for a specific word in different Moments, because the embedding feature vector is same, the output global weight does not change.

In the end, the final weight of a word in a Moment is defined as:

$$g_f = g_l \cdot g_g \tag{5}$$

Compared with other possible definitions, such as computing the mean value of global and local weight as the final weight, this definition achieves best performance.

3.3 Semi-supervised Image Classification

As we stated in Section 3.2, we classify all Moment images into n different categories as the basis of image-guide attention mechanism. Considering that we do not have any label for the Moment images, we classify each image by extracting and clustering its deep neural network features. The whole process is shown as follows. First, we extract a deep-level 2048-dimensional feature vector for each image from the last "pool5" layer of ResNet-50 proposed by He et al [6]. After that, we cluster these feature vectors by k-means clustering. We determine the value of k based on the well-known Silhouette Coefficient. To reduce the time complexity, we replace the mean distance of a sample to all samples of a cluster with the distance between this sample and the centroid of this cluster. We set k from 10 to 100 and find that when k is larger than 60, there is a marked decline for the Silhoette Coefficient. Therefore we set k = 60 and obtain 60 categories with their corresponding Moment images. Next, we manually combine several categories (which makes it semi-supervised) that we judge to be the same category, generate 50 categories in the end and label them according to their corresponding images. The names of the 50 categories are shown in Table 1. An image of test set is classified as the category whose centroid holds the minimum Euclidean distance with the image's high-level feature. To evaluate the performance of the image classification method, for each category, we randomly sample 500 images and ask two volunteers to judge whether it is accurate to classify an image into this category. The average accuracy for all categories is 88.7%, with a standard deviation of 8.89, while 43 of 50 categories are higher than 80%. Several typical categories' classification results are shown in Figure 4. We can see that the semi-supervised image classification has an adequately high accuracy. Noticing that our categories have some different characteristics from other well-known social networking services, such as Pinterest[2], which defines 34 available categories for users to choose from. Since our goal to generate categories is to improve the model's performance, so the exact definitions of the categories are less critical.

Figure 4: Several typical categories' clustering results. (a) Flower (b) Meal (c) Cosmetics (d) Chat Screenshot.

3.4 Architecture of BiATT-LSTM

We formally describe the architecture of the BiATT-LSTM model shown in Figure 3. The basic LSTM contains a word embedding layer, an LSTM layer, two fully-connected layers with ReLU nonlinearity function and a softmax layer which outputs two values that represent probabilities of WeChat Business Moment and WeChat non-business Moment. For the rest of the model, the word embedding feature is not only added as an input term of the LSTM layer,

[2]https://www.pinterest.com/

Table 1: Name of each category

Indoor Selfie	Snack	Cosmetic Tips
Pet	Landscape Photo	Display Rack
Bed	Tourist Photo	Hand & Leg
Big Word Ad	Sunglass & Handbag	Wallet & Accessory
Small Group Photo	Photoshop Photo	Fruit & Cake
Poster	Star	WeChat Moment
Chart	Beauty Ad	Motto
Pink Goods	Holding Something Selfie	WeChat Expression
Child	Necklace & Bracelet	QR-code
Flower	Baby	WeChat Wallet
Cosmetics	Full-Length Photo	Chat Screenshot
Cosmetic Ad	Special Effects Photo	Other Ad
Activity	Very Long Picture	Comic
Large Group Photo	Meal	Essay
Building	Outdoor Selfie	Other Goods
TV & Poster Screenshot	Face Mask Selfie	Shoes
Toy	Clothes	

"Tourist Photo" represents full-length photo of tourist in a tourism scene.
"Photoshop Photo" represents photo with words/graphs using photoshop.

but also as the input to learn the word's global weight. On the other hand, for each word we compute the 50-dimensional category correlation word feature that represents the correlation between the word and image categories, and thus figure out the local weight of the word in a specific Moment via the 50-dimensional category distribution feature. In the end, we feed the bilateral weights of a word into the LSTM by modifying the equations representing the operations of the LSTM cell [4] as follows:

$$\begin{aligned} i_t &= \sigma(g_f W_{xi} x_t + W_{hi} h_{t-1} + b_i) \\ f_t &= \sigma(g_f W_{xf} x_t + W_{hf} h_{t-1} + b_f) \\ o_t &= \sigma(g_f W_{xo} x_t + W_{ho} h_{t-1} + b_o) \\ g_t &= \phi(g_f W_{xc} x_t + W_{hc} h_{t-1} + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \phi(c_t) \end{aligned} \tag{6}$$

Where for the current focusing word at time step t, x_t are the 32-dimensional word embedding feature and, g_f is the word's final weight computed by its global and local weights. W and b are learning parameters of the LSTM network.

Finally, we fuse the hidden state of last time with the category distribution feature of the Moment and input them to the fully-connected layers. On one hand, we notice that in addition to reducing the network complexity, inputting the hidden state of last time to the next fully-connected layer leads to better performance than inputting the hidden states of all times. Meanwhile, we find that for the utility of image information in a Moment, even though the image-guide attention mechanism achieves better performance than late fusion, but fusion mechanism could still provide complementary information and improve the performance. Therefore, we still add the category distribution feature as part of the input of the last fully-connected layers. Cross-entropy loss is employed as the loss function and a mini-batch gradient descent algorithm with an adaptive learning rate is used to optimize the network.

4 EXPERIMENTS

To demonstrate the effectiveness of our BiATT-LSTM for the task, we perform a two-level "vertical" and "parallel" experiments that show the results of different models on several measures. For "vertical" experiments, we compare the basic LSTM framework with other three popular frameworks for text classification, text-based decision tree framework, Doc2vec-based framework [7] and Latent Dirichlet Allocation-based framework [3] to show its effectiveness. Then we feed bilateral-attention mechanism into model as shown in Figure 3, and compare the experiment results. For "parallel" experiments, intrinsically, we consider the BiATT-LSTM as a new framework that jointly models vision and language content in a novel way, we thus compare our model's result with other multi-modality frameworks. In recent years, a number of innovative models [1, 2, 8, 18, 19, 21] implement different tasks based on jointly modeling image and language content. However, with the restrictions of 1) our task is a classification task, 2) one text corresponds to multiple images and 3) the language style are informal, some of them are not suitable for our work. Therefore, we compare our methods with four models, the normal late fusion model, factorization machines [12], Aishwarya's deeper LSTM Q + norm I model (LSTM Q) [1] and You's Cross-modality Consistent Regression model (CCR) [21]. You's CCR model assumes that different modalities should be consistent in terms of depicting the same subject, it thus imposes consistent constraints across related but different modalities (visual and textual). Aishwarya's LSTM Q model replaces the concatenation of visual and textual feature with common space mapping and element-wise multiplication. Factorization machines could model all interactions between features using factorized parameters, thus they are able to estimate interactions even in problems with huge sparsity. It adapts our category distribution feature which is a sparse vector. All baseline models are suitable for joint visual-textual classification task.

4.1 Dataset

We collect a dataset that consists of 570 users with their 37,359 WeChat Moments and 109,545 Moment images, from Mar 21, 2016 to July 21, 2016. All of these users are VIPs of a cosmetics brand. We collect data from this kind of users because a considerable amount of their Moments is about WeChat Business. In addition to cosmetics, they also advertise other products such as clothes, shoes, restaurants, start-ups, exotic fruits, luxuries, platforms, laundry detergents, high-tech gears, and so on. To train and test the WeChat Business Moment classifier, two researchers randomly select 10078 Moments and respectively label part of them in two categories, WeChat Business Moment and WeChat non-business Moment. The similar label proportion of WeChat Business Moment (nearly 43%) by both researchers and a thorough process of sub-sample validation ensures the consistency of the labeling process. In the end, 4309 of 10078 Moments are positive samples that are related to WeChat Business. For all experiments, We randomly select 80% Moments as train samples and 20% Moments as test samples, and cross-validated model hyperparameters based on random 10% samples of the train set.

4.2 Experimental Settings

For vertical experiments, we first demonstrate the effectiveness of LSTM. For text-based decision tree, we extract text feature by TF-IDF [16] and directly train a decision tree for classification, for Doc2vec-based and LDA-based frameworks, we respectively extract the text feature by Doc2vec and the combination of TF-IDF and LDA, and input the feature to Multi-layer Perceptron for classification. We set both text feature dimension of Doc2vec and topic number of LDA as 300 for best performance. After that, we compare the BiATT-LSTM framework with the baseline LSTM. Consistent with [21], four measures, accuracy, precision, recall and F-measure are used to measure the performance.

For the parallel experiment, we compare the model's performance with late fusion, factorization machines, LSTM Q and CCR. For CCR, as [21] does, we add a new loss term that imposes consistent constraints across visual and textual modalities. For factorization machines, we replace the top fully-connected layers with factorization machines layers. For LSTM Q, we normalize the visual feature (CD feature) as [1], and also implement common space mapping and element-wise multiplication.

On the other hand, considering the sample number for our dataset, we also focus on a simulation of using a large dataset for our task, and prove that bilateral-attention mechanism still have strong capacity to improve the performance. Specifically, because of the complicate language environment, very large word vocabulary and relatively limited sample number of the dataset, we find that a great amount of words in test set only exist few times or even not exist in training set. This situation leads to inaccurate word embedding for these words and make the model's performance far from using a large dataset, from which, the model can learn accurate word embedding for each word of test set. For our task, pre-trained word embedding is difficult to be transfered and used since 1) it's difficult to find large dataset that contains all these words/expressions to train the model 2) most importantly, the task is high-level, even words with completely different semantic meanings (e.g. "washing powder" and "face mask") in ordinary datasets can have similar attributes toward our task (positive to WeChat Business), which weakens the significance of pre-trained word embedding. Therefore, to simulate the situation of using a large dataset, we manually incorporate a strong word feature into the LSTM cell. In particular, for a word and a label, we construct a correlation coefficient between them in the same way as constructing the correlation between word and category, but also adding the test samples of dataset instead of just using training samples, this operation artificially makes up for the bad word embedding quality of a test word which only exists few times and simulate a more close situation of using large dataset. Since there are two types of labels for the text of our binary classification task, the word feature is a 2-dimension feature vector where each node records the correlation coefficient between this word and a specific label. In our experiments, we also show the performance of different models based on this simulation.

4.3 Experimental Results

Table 2 shows the vertical experiment results, we can see that using image category distribution feature, the accuracy reaches 77.12%.

Table 2: Vertical comparison of accuracy, precision, recall and F-measure based on different frameworks. "LSTM(S)" represents the experiments of simulating large dataset.

	Accuracy	Precision	Recall	F-measure
Category Distribution Feature	77.12	76.86	67.77	71.96
DecisionTree	80.41	78.09	76.27	77.17
Doc2vec[7]	81.23	83.16	71.18	76.70
TFIDF+LDA[3]	77.71	74.80	73.39	74.09
LSTM	87.34	86.18	84.37	85.27
BiATT-LSTM	89.85	91.48	84.48	87.84
LSTM(S)	93.26	89.85	95.23	92.46
BiATT-LSTM(S)	96.01	96.05	94.67	95.13

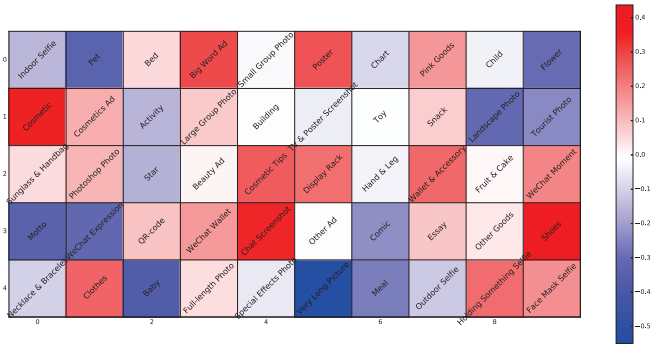

Figure 5: Correlation between WeChat Business and each image category.

A logistic regression model is implemented to compute the correlation between WeChat Business and each image category. Figure 5 shows the logistic regression coefficient of each category, we can see that most of image categories have the capacity (strong positive or negative correlation) to classify the Moments, but only using them as direct features for our task cannot exert their stronger potential as a guidance to adjust the text word weight. For text information, using LSTM model could reach 87% accuracy for the task, it is better than LDA-based and Doc2vec-based framework, which demonstrates that LSTM achieves better performance for our task. Most importantly, bilateral-attention mechanism remarkably improves the performance for both experiments with common setting and setting of simulating large dataset, it respectively improve the accuracy from 87.34% to 89.85% and from 93.26% to 96.01%, which is significant considering that the baseline accuracy is high. In addition to the highest accuracy, BiATT-LSTM achieves a balanced performance for both precision and recall, which indicates that most of the Moments it predicts as positive are indeed WeChat Business Moment, and most of the real WeChat Business Moments are identified.

In the parallel experiment, from Table 3, it can be seen that the bilateral-attention mechanism, as a new approach to jointly modeling visual and textual modalities, helps the network learn better for the visual-textual interaction and achieves excellent performance on different experiment settings, it outperforms other approaches on almost all situations. Also, we could see that a combination of bilateral-attention mechanism and late fusion can still improve the

performance, which demonstrates that bilateral-attention mechanism is compatible with other frameworks, because it acts on the bottom layers of the model, which makes it potential to coexist with other approaches that act on top layers of the model.

Table 3: Parallel comparison of accuracy, precision, recall and F-measure based on different frameworks. "CD" represents category distribution feature, "LF" represents normal late fusion.

	Accuracy	Precision	Recall	F-measure
LSTM+CD+LF	88.98	90.02	84.04	86.93
LSTM+CD+CCR[21]	89.18	90.17	84.27	87.12
LSTM+CD+FM[12]	89.42	89.90	85.22	87.50
LSTM+CD+LSTM Q[1]	88.65	91.53	81.39	86.17
BiATT-LSTM	89.85	91.48	84.48	87.84
BiATT-LSTM+CD+LF	90.33	91.36	85.70	88.49
LSTM(S)	93.26	89.85	95.23	92.46
LSTM(S)+CD+LF	94.27	97.57	89.02	93.10
LSTM(S)+CD+CCR	94.63	92.48	95.35	93.89
LSTM(S)+CD+FM	94.94	94.81	93.55	94.18
LSTM(S)+CD+LSTM Q	94.13	94.02	92.36	93.81
BiATT-LSTM(S)	96.01	96.05	94.67	95.13
BiATT-LSTM(S)+CD+LF	96.20	95.47	95.78	95.63

Table 4: Typical words with their local weights in a Moment that contains images of a single following category. Sigmoid unit is replaced with ReLU for a better display of words' differences. Values with bold format are relatively high weights for each word.

	Cosmetic Ad	Meal	Landscape Photo	Shoes	Motto	Chat Screenshot
Skin	**3.56**	0	0	0	0	0.11
Taste	0.26	**0.63**	0	0	0	0.28
Scenery	0	0.21	**1.64**	0	0.25	0
Color	**0.48**	0	0.03	**0.47**	0.16	0.09
Sunshine	0	0.21	**0.53**	0	**0.44**	0
Feedback	1.85	0	0	0.80	0	**3.52**

Table 4 shows typical words' local weights in a Moment that contains images of a specific category, we replace sigmoid unit with ReLU which can show the difference of words in a better way. We could see that "skin", "taste", "scenery", "color", "sunshine" and "feedback" respectively gain high local weight for categories of "Cosmetic Ad", "Meal", "Landscape Photo", "Shoes", "Motto" and "Chat Screenshot". Figure 6 illustrates how bilateral-attention mechanism works. In (a), we test two artificial Moments with the same text but different images. For the global weights of the words, the words that could provide significant information to judge whether a Moment is related to WeChat Business are learned with high global weights. For example, "along the trip", "mountain", "whitening", "face masks" possess high global weights, thus the first two words are marked words for WeChat non-business Moment and the last two words are the counterpart. For Moment *A*, with several images belonging to Cosmetic and Cosmetic Ad categories, "whitening" and "face masks" hold high local weights and "along the trip", "mountain" hold low local weights. The final word weights have an clear tendency to strengthen the words in the latter part of the sentence. On the other hand, for Moment *B*, with the image environment related to travel, the words in the latter part of the sentence are weakened with low local weights. In the end, with totally different distributions of final word weights for the same sentence, Moment *A* will be predicted as positive and Moment *B* will be predicted as negative. In (b), it tests two Moments with the same images and different texts, still ,the bilateral attention mechanism could adjust the weights as we would expect and identify Moment *A* as a WeChat Business Moment.

5 CONCLUSIONS

In this paper, we propose a Bilateral-Attention task driven LSTM network to identify WeChat Business related Moments. To make better use of image information of a Moment, we incorporate an image-guide attention mechanism to automatically learn a word's local weight on its corresponding specific Moment. On the other hand, to extract the overall importance of each word for the classification task and strengthen the significant words, a self-learned attention mechanism is implemented to learn words' global weight. We figure out the final weight and feed it into the LSTM cells to adjust the final importance of a word in a text. Two-level experiments demonstrate that BiATT-LSTM remarkably improves the model's performance on all measures. In particular, the image-guided attention mechanism provides a novel approach to make the best use of the weaker visual modality for joint visual-textual learning tasks.

REFERENCES

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2017. VQA: Visual Question Answering. *International Journal of Computer Vision* 123, 1 (2017), 4–31.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[4] Alex Graves et al. 2012. *Supervised sequence labelling with recurrent neural networks*. Vol. 385. Springer.

[5] Zhao Guo, Lianli Gao, Jingkuan Song, Xing Xu, Jie Shao, and Heng Tao Shen. 2016. Attention-based LSTM with Semantic Consistency for Videos Captioning. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 357–361.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv preprint arXiv:1512.03385* (2015).

[7] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 1188–1196.

[8] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. 2017. Person Search with Natural Language Description. *arXiv:1702.05729* (2017).

[9] Zhuqi Li, Lin Chen, Yichong Bai, Kaigui Bian, and Pan Zhou. 2016. On Diffusion-restricted Social Network: A Measurement Study of WeChat Moments. *IEEE International Conference on Communications* (2016).

[10] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[11] Jiezhong Qiu, Yixuan Li, Jie Tang, Zheng Lu, Hao Ye, Bo Chen, Qiang Yang, and John E. Hopcroft. 2016. The Lifecycle and Cascade of WeChat Social Messaging Groups. *Proceedings of. ACM International Conference on World Wide Web (WWW) Pages 311-320* (2016).

[12] Steffen Rendle. 2010. Factorization Machines. In *ICDM 2010, the IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December*. 995–1000.

[13] Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. (2016), 4613–4621.

[14] Kunal Swani, Brian P. Brown, and George R. Milne. 2014. Should tweets differ for B2B and B2C? An analysis of Fortune 500 companies' Twitter communications. *Industrial Marketing Management* 43, 5 (2014), 873–881.

[15] Yang Wang, Yao Li, Bryan Semaan, and Jian Tang. 2016. Space Collapse: Reinforcing, Reconfiguring and Enhancing Chinese Social Practices through WeChat. In *ICWSM*.

[16] Ho Chung Wu, Robert Wing Pong Luk, Kam Fai Wong, and Kui Lam Kwok. 2008. Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems (TOIS)* 26, 3 (2008), 13.

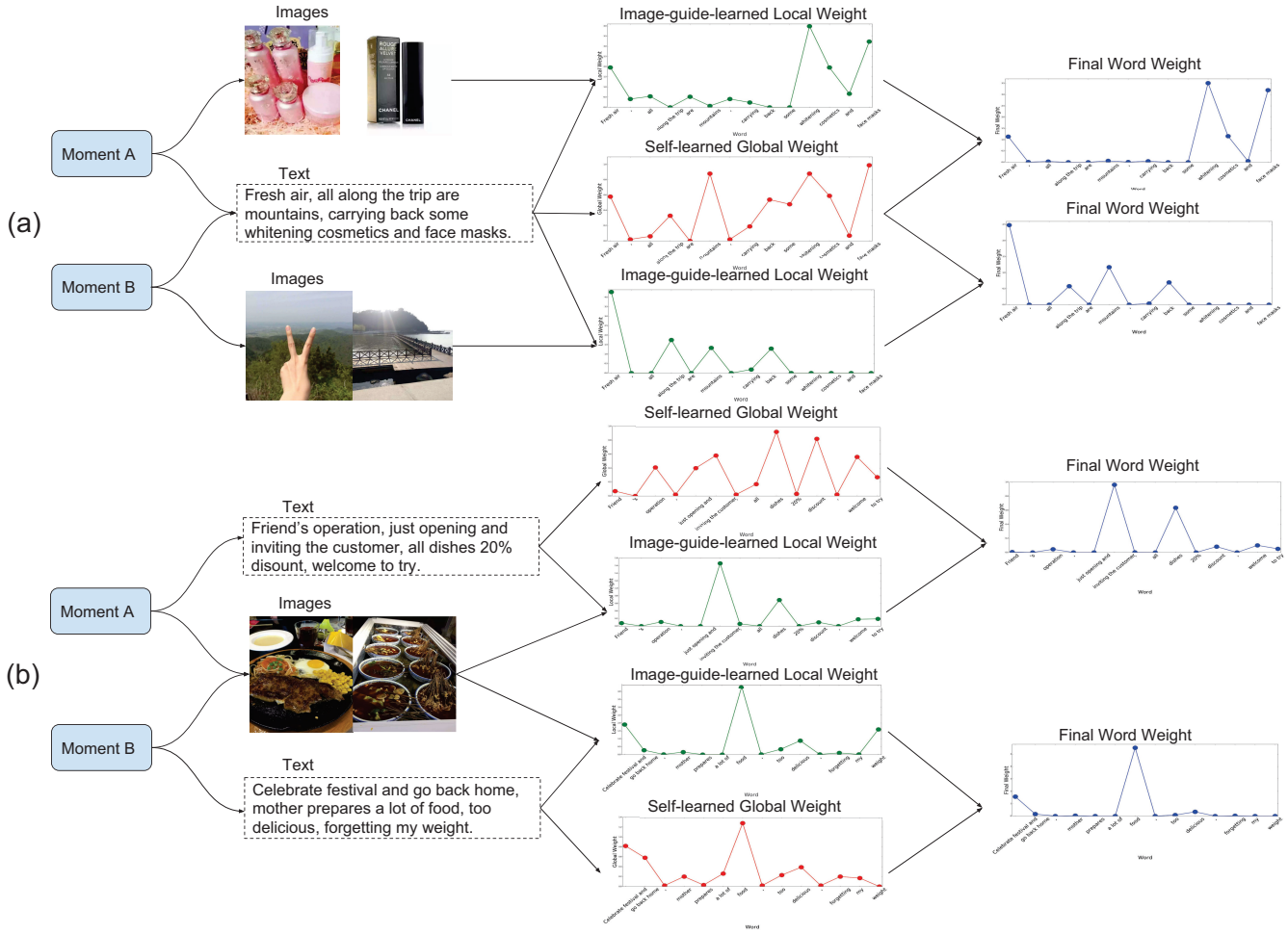

Figure 6: Visualization of the bilateral attention mechanism on two artificially designed examples (The word segmentation is based on Chinese, each Chinese word may correspond to multiple English words). (a) represents two Moments with the same text but different images. (b) represents two Moments with the same images but different texts.

[17] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. 14 (2015), 77–81.

[18] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. (2016), 21–29.

[19] Quanzeng You, Liangliang Cao, Hailin Jin, and Jiebo Luo. 2016. Robust Visual-Textual Sentiment Analysis: When Attention meets Tree-structured Recursive Neural Networks. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1008–1017.

[20] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. (2016), 4651–4659.

[21] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia. In *ACM International Conference on Web Search and Data Ming (WSDM)*. 13–22.

[22] Chengxi Zang, Peng Cui, and Christos Faloutsos. 2016. Beyond Sigmoids: The NetTide Model for Social Network Growth, and Its Applications. In *The ACM SIGKDD International Conference*. 2015–2024.

[23] Shuangfei Zhai, Keng Hao Chang, Ruofei Zhang, and Zhongfei Mark Zhang. 2016. DeepIntent: Learning Attentions for Online Advertising with Recurrent Neural Networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1295–1304.