Capsule Network with Interactive Attention for Aspect-Level Sentiment Classification

[12]Chunning Du, [12]Haifeng Sun, [12]Jingyu Wang*, [12]Qi Qi, [12]Jianxin Liao
[12]Tong Xu, [12]Ming Liu
[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]EBUPT Information Technology Co., Ltd., Beijing 100191, China
{duchunning,sunhaifeng_1,wangjingyu,qiqi}@ebupt.com

Abstract

Aspect-level sentiment classification is a crucial task for sentiment analysis, which aims to identify the sentiment polarities of specific targets in their context. The main challenge comes from multi-aspect sentences, which express multiple sentiment polarities towards different targets, resulting in overlapped feature representation. However, most existing neural models tend to utilize static pooling operation or attention mechanism to identify sentimental words, which therefore insufficient for dealing with overlapped features. To solve this problem, we propose to utilize capsule network to construct vector-based feature representation and cluster features by an EM routing algorithm. Furthermore, interactive attention mechanism is introduced in the capsule routing procedure to model the semantic relationship between aspect terms and context. The iterative routing also enables encoding sentence from a global perspective. Experimental results on three datasets show that our proposed model achieves state-of-the-art performance.

1 Introduction

Aspect-level sentiment classification is a fine-grained task in the field of sentiment analysis (Pang and Lee, 2008; Liu, 2012), which aims to infer the sentiment polarity (e.g., positive, neutral, negative) of a sentence with respect to the aspect. It demands to differentiate sentiments towards different targets when there are multiple targets in one sentence. For example, given the mentioned aspect term {*food, price, drinks*}, and the sentence is *"The food was definitely good, but when all was said and done, I just could not justify it for the price including 2 drinks, $100/person."* For aspect term *food*, the sentimental polarity is positive, but for aspect term *price*, the polarity is negative while for aspect term *drinks*, the polarity is neutral. Recently, with the development of deep learning techniques, various neural networks are designed for this task and obtain promising results (Wang et al., 2016a; Ma et al., 2017; Fan et al., 2018).

The main challenge in aspect-level sentiment classification is that one sentence expresses multiple sentiment polarities, resulting in overlapped feature representation. Take the same example above, the sentence simultaneously reviews on 'food', 'price', and 'drinks', and expresses three different sentiment polarities. The highly overlapped features will confuse the classifier seriously. However, most existing methods only keep the most active feature by max-pooling operation or utilize attention mechanism to find the sentimental words, which fails to distinguish the overlapped features.

Therefore, we propose a novel capsule network and iterative EM routing method with interactive attention (IACapsNet) to solve this problem. Capsule network (Hinton et al., 2011; Sabour et al., 2017; Hinton et al., 2018) constructs vector-based feature representation. Capsules in adjacent layers are connected by dynamic routing, which shows strengths in distinguishing overlapped features by feature clustering (Sabour et al., 2017; Zhang et al., 2019). In the aspect-level sentimental classification task, the vector-based overlapped sentimental features towards different aspect terms will be clustered by an Expectation-Maximization (EM) routing algorithm, which makes the subsequent classification more clear. Furthermore, we further devise an interactive attention-based routing mechanism in order to highlight the word-level difference and model the semantic relationship between aspect terms and context.

Moreover, our iterative routing mechanism can

*Corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 5489–5498,
Hong Kong, China, November 3–7, 2019.

be viewed as a top-down attention mechanism, which is more efficient because of the global perspective compared to the standard attention mechanism. Standard attention mechanism in this task only considers a part of the context information in a sentence without considering the overall meaning conveyed by the sentence, which may introduce noise and downgrade the prediction accuracy, especially for complex sentences. For example, in an ironic statement for the aspect term "mac os": "*Maybe the mac os improvement were not the product they want to offer*.", the standard attention mechanism will highlight the sentimental word '*improvement*' and confuse the classifier to make the wrong prediction to be positive. Our routing mechanism can tackle this by adjusting the contribution of each low-level capsule based on the high-level capsules (overall representation), and the iterative update makes the overall representation more accurate compared to other similar top-down attention (Liu et al., 2018; Zhao and Zhang, 2018).

Our proposed model (IACapsNet) is evaluated on three datasets: laptop, restaurant datasets from the SemEval 2014 Task 4 and Twitter collection. The experimental results show that our model outperforms other baseline methods and achieves state-of-the-art performance. Our contributions are summarized as follows:

- We apply capsule network to aspect-level sentiment classification to tackle the overlapped features by feature clustering. To the best of our knowledge, there is no work that investigates the performance of capsule network in this task.
- An interactive attention mechanism is introduced in the capsule routing to help model the semantic relationship between aspect term and context.

2 Related work

2.1 Aspect Level Sentiment Classification

Traditional approaches have designed rich features about content and syntactic structures to capture the sentiment polarity (Jiang et al., 2011; Pérez-Rosas et al., 2012). However, These feature-based methods are labor-intensive and the performance highly depends on the quality of the features. Recently, deep learning methods are becoming popular for aspect-level sentiment classification. Recurrent Neural Networks (RNNs) are the most commonly used technique for this task (Tang et al., 2016a). The attention mechanism is further introduced to model the target-context association (Wang et al., 2016b; Li et al., 2017; Ma et al., 2017). Furthermore, Fan et al. (2018) proposed MGAN to integrate fine-grained attention mechanisms, which is employed to characterize the word-level interactions between aspect and context words. Very recently, CNN-based models have shown the strengths in efficiency to tackle the aspect-level sentiment classification (Xue and Li, 2018; Huang and Carley, 2018; Li et al., 2018). However, all the previous methods utilize static pooling operation or attention mechanism to locate the sentimental words, which fails to handle the overlapped features. We introduces vector-based feature representation and feature clustering to address this.

2.2 Capsule Network

Capsule network was proposed to improve the representational limitations of CNN and RNN by extracting features in the form of vectors. The technique was firstly proposed in (Hinton et al., 2011) and improved in (Sabour et al., 2017; Hinton et al., 2018), which is mainly devised for image processing domain. Introducing capsules allows us to utilize a routing mechanism instead of pooling operation to generate high-level features which is a more efficient way for features encoding. Routing-by-agreement is able to cluster features in an iterative way, which achieved impressive performance recognizing highly overlapped digits.

Several types of capsule networks have been proposed for natural language processing. Yang et al. (2018) investigated capsule networks for text classification. They also found that capsule networks exhibit significant improvement when transferring single-label to multi-label text classification. Similar property has also been observed in the task of relation extraction (Zhang et al., 2019, 2018). However, interactive word-level attention is not considered in these typical capsule routing methods.

3 Model

In this section, we describe the proposed capsule network with interactive attention (IACapsNet) in details. The aim of aspect-level sentiment classification is to predict the sentiment class y of a

sentence over a specific aspect term, where $y \in \{positive, negative, neutral\}$. The overall architecture is shown in Figure 1. It consists of the input embedding layer, bidirectional RNN layer, primary capsule layer, and output layer.

3.1 Input Embedding Layer

The context's input representations of IACapsNet include word embeddings w_n and position embeddings p_n. The aspect term's input representation only consists of word embedding w_n^a.

Word embedding is a distributed representation of a word, where words from the vocabulary are mapped to vectors. Initializing words vectors via pre-trained word vectors can improve the performance due to their ability to capture syntactic and semantic information of words from large scale unlabeled text. In our model, we employ the pre-trained word vector GloVe (Pennington et al., 2014) to obtain the fixed word embedding $w_n, w_n^a \in \mathbb{R}^{d_w}$, where d_w is the word vector dimension.

Considering that the context words with closer distance to an aspect may have higher influence on the sentiment analysis, we introduce position embedding to encode the relative distance r_n from word w_n to the aspect term. We define the position embedding matrix $P \in \mathbb{R}^{d_p \times N}$, which is randomly initialized and updated during the training process. Here, d_p is the position embedding dimension and N denotes the length of the sentence. The corresponding word's position embeddings p_n can be obtained by looking up the position embedding matrix P using r_n.

The input representation for each context word is the concatenation of word embeddings and position embeddings: $x_n = [w_n; p_n] \in \mathbb{R}^{d_w + d_p}$.

3.2 Bidirectional Recurrent Networks Layer

The recurrent neural networks can capture long-distance dependencies within a sentence. A bidirectional recurrent neural network is the first layer of IACapsNet. The forward direction captures the left context h_l for a word and the backward direction captures the right context h_r. We concatenate the left context and the right context as the contextualized word representation h_n^c, $h_n^a \in \mathbb{R}^{2 \times d_l}$, where d_l is the dimension of hidden state, h^a and h^c are the word representations for aspect term and context, respectively.

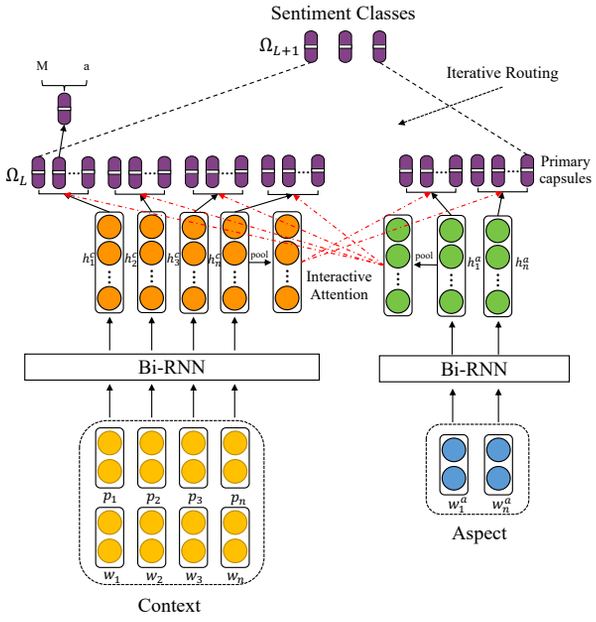

Figure 1: The architecture of IACapsNet

3.3 Primary Capsule Layer

The primary capsule is a group of neurons obtained from the output of the convolutional operation performed on h_n^a and h_n^c. So, the output of capsule is a vector representing different properties of the same objective. In aspect-level sentiment classification task, the properties may contain the sentiment and aspect term features.

EM-based routing method (Hinton et al., 2018) is implemented in our model, and except for the high-dimensional output M, there is one more activation probability $\boldsymbol{a}$ in our capsule, which is like the activity in a standard neural net (shown in Figure 1).

3.4 Interactive Attention EM Routing

We have already decided on the outputs of all the capsules Ω_L in primary capsule layer and we now want to decide which capsules Ω_{L+1} to active in the layer above and how to assign each active low-level capsule to one active higher-level capsule.

The vector-based features get clustered in the high-level capsules by an EM based algorithm where the outputs of high-level capsules play the role of Gaussians and the output vectors of low-level capsules play the role of the datapoints. The means, variances, and activation probabilities of the output capsules, as well as the assignment probabilities R of the input capsules are iteratively updated by alternating between an E-step and an M-step. It can also be viewed as a parallel atten-

tion mechanism in the opposite direction, which can adjust the low-level word's contribution based on the sentence global representation. Moreover, in order to model the semantic relationship between aspect term and context, we further devise an interactive attention-based routing mechanism.

3.4.1 M-step

The M-step holds the assignment probabilities R constant and adjusts each Gaussian (i.e. high-level capsules) to maximize the sum of the weighted log probabilities that the Gaussian would generate the datapoints (i.e. low-level capsules) assigned to it. This procedure aims to obtain the overall representation among the low-level capsules for a given iteration.

Firstly, every primary capsule i is transformed by W_{ij} to cast a vote $V_{ij} = M_i W_{ij}$ for the output of high-level capsule j. And we can get the mean μ_j of the votes from the input capsules and the variance σ_j about that mean for each dimension h:

$$\mu_j^h = \frac{\sum_i R_{ij} V_{ij}^h}{\sum_i R_{ij}}, \tag{1}$$

$$(\sigma_j^h)^2 = \frac{\sum_i R_{ij} (V_{ij}^h - \mu_j^h)^2}{\sum_i R_{ij}}, \tag{2}$$

where μ_j^h is the h^{th} component of the capsule j's vectorized output M_j.

The activation probability of capsule j is calculated by

$$cost_j^h = (\beta_u + log(\sigma_j^h)) \sum_i R_{ij}, \tag{3}$$

$$a_j = \text{sigmoid}(\lambda(\beta_\alpha - \sum_h cost_j^h)), \tag{4}$$

where β_u and β_α are trainable parameters denoting fixed cost per input capsule when not activating it and fixed cost for coding the mean and variance of capsule j when activating it. The variance σ reflects the degree of agreement. An intuitive understanding of this activation probability is that if the votes from low-level capsules are not agreed on one high-level capsule, the activation of the high-level capsule should be low. λ is an inverse temperature parameter set $1e{-}3$ with a fixed schedule.

3.4.2 E-step

The E-step adjusts the assignment probabilities R for each datapoint (i.e. low-level capsule) to the Gaussian (i.e. high-level capsules). This procedure aims to adjust the contribution of each capsule based on the high-level capsule (i.e. overall representation) for a given iteration.

We firstly compute the negative log probability density of the vectorized vote under the j's Gaussian distribution:

$$p_j = \frac{1}{\sqrt{\prod_h^H 2\pi(\sigma_j^h)^2}} \exp(-\sum_h^H \frac{(V_{ij}^h - \mu_j^h)^2}{2(\sigma_j^h)^2}). \tag{5}$$

For capsule i in primary capsule layer, the assignment probability is adjusted by:

$$R_{ij} = \frac{a_j p_j}{\sum_j a_j p_j}. \tag{6}$$

Alternating E-step and M-step will route the output of capsule to a capsule in the layer above that receives a cluster of high-dimensional features.

Algorithm 1 Interactive attention EM routing. Capsule i and j denote a low-level and high-level capsule. Ω_L and Ω_{L+1} denote the low-level and high-level capsules set, respectively.

procedure INTERACTIVE ATTENTION EM ROUTING($a_i, V_{ij}, \alpha^{a\to c}, \alpha^{c\to a}$)
  $\forall$ capsules i and capsules j: $R_{ij} = \frac{1}{|\Omega_{L+1}|}$
  $\forall$ capsules i from context: $a_i = a_i * \alpha_n^{a\to c}$
  $\forall$ capsules i from aspect: $a_i = a_i * \alpha_n^{c\to a}$
  n is the token index from which the capsule i comes.
  for r iterations **do**
    $\forall j \in \Omega_{L+1}$: M-STEP($a_i$, R_{ij},V_{ij},j)
    $\forall i \in \Omega_L$: E-STEP(μ_j, σ_j,a_j,V_{ij},i)
  end for
  return a_j
end procedure
procedure M-STEP(a_i,R_{ij},V_{ij})
  $\forall$ capsule i: $R_{ij} = R_{ij} * a_i$
  $\forall$ capsule j: compute μ_j and σ_j by Eq. 1 and 2
  $\forall$ capsule j: compute a_j by Eq. 4
end procedure
procedure E-STEP(μ_j,σ_j,a_j,V_{ij})
  $\forall$ capsule j: compute p_j and update R_{ij} by Eq. 5 and 6
end procedure

3.4.3 Interactive Attention

The overlapped features in primary capsule layer can be routed and clustered to high-level capsules in an iterative way. However, purely alternating the E-step and M-step will ignore the relationship between the context and aspect term. It has been demonstrated that target and context can determine the representation of each other. And the coordination of targets and their contexts can remarkably enhance the performance of sentiment classification (Ma et al., 2017; Fan et al., 2018). Word-level attention, which has already been demonstrated to be essential is also ignored. So, apart from the assignment probabilities, we further introduce an interactive attention weight α which is learned interactively between the context and the aspect term.

Specifically, we implement scaled dot-product attention which can be described as mapping a query and a set of key-value pairs to a weight on the word-level token. The queries are the averaged representation of the context h^c and the aspect term h^a which are transformed to dimension d_k by trainable parameters:

$$q_c = \frac{\sum_n^N h_n^c}{N} * W_c^q \in \mathbb{R}^{d_k}, \tag{7}$$

$$q_a = \frac{\sum_n^N h_n^a}{N} * W_a^q \in \mathbb{R}^{d_k}, \tag{8}$$

The keys are the corresponding words' representation from aspect term and context which are also transformed to dimension d_k:

$$k_n^c = h_n^c * W_c^k \in \mathbb{R}^{d_k}, \tag{9}$$

$$k_n^a = h_n^a * W_a^k \in \mathbb{R}^{d_k}, \tag{10}$$

where $W_c^q, W_a^q, W_c^k, W_a^k \in \mathbb{R}^{2d_l * d_k}$

The attention weights can be computed as follows:

$$s_n^{a\to c}(q_a, k_n^c) = \frac{k_n^c * q_a^\mathsf{T}}{\sqrt{d_k}}, \tag{11}$$

$$s_n^{c\to a}(q_c, k_n^a) = \frac{k_n^a * q_c^\mathsf{T}}{\sqrt{d_k}}, \tag{12}$$

$$\alpha_n^{a\to c} = \frac{\exp(s_n^{a\to c}(q_a, k_n^c))}{\sum_{n=1}^N \exp(s_n^{a\to c}(q_a, k_n^c))}, \tag{13}$$

$$\alpha_n^{c\to a} = \frac{\exp(s_n^{c\to a}(q_c, k_n^a))}{\sum_{n=1}^N \exp(s_n^{c\to a}(q_c, k_n^a))}. \tag{14}$$

In order to ensure a proper magnitude of s_n avoiding pushing the softmax function into regions where it has extremely small gradients, we introduce the scaling factor $\frac{1}{\sqrt{d_k}}$ following (Vaswani et al., 2017). We can thus get the word-level significance for each token in the context and aspect term in an interactive way, which will be adopted on each primary capsule's activation probability a. We detail the whole routing algorithm in Algorithm 1.

3.5 Training Objective

In IACapsNet, each top-level capsule corresponds to a sentiment category. The activation probability a of each top-level capsule represents the probability that the input sentence belongs to the corresponding category. We use a spread margin loss, L_k for each top-level capsule k to directly maximize the gap between the activation of the target class (a_t) and the activation of the other classes. The total loss L is simply the sum of the losses of all top-level capsules:

$$L = \sum_{k\neq t} (max(0, m - (a_t - a_k))^2, \tag{15}$$

where m is the margin which we set 0.9 with a fixed schedule.

4 Experiments

In this section, we conduct extensive experiments on three datasets to evaluate the proposed IACapsNet.

4.1 Experimental Setup

The experiments are implemented on three datasets. The first two datasets are from the SemEval 2014 Task 4 (Pontiki et al., 2014), which contains reviews about laptops and restaurants, respectively. The third one is a Twitter dataset collected by (Dong et al., 2014). The statistics of these datasets are listed in Table 2. Following (Tang et al., 2016c), conflict category is removed from the SemEval 2014 datasets to avoid datasets getting unbalanced. Sentences are zero-padded to the length of the longest sentence in respective dataset. Results are measured by accuracy and Macro-Averaged F1 score.

In our experiments, the pre-trained GloVe (Pennington et al., 2014) is used to initialize the word embeddings from context and aspect term. The

Model	Laptop		Restaurant		Twitter	
	ACC	Macro-F1	ACC	Macro-F1	ACC	Macro-F1
ATAE-LSTM (Wang et al., 2016a)	69.27	-	78.50	-	69.88	-
TD-LSTM (Tang et al., 2016b)	68.83	68.43	78.00	66.73	66.62	64.01
MemNet (Tang et al., 2016c)	72.37	-	80.32	-	68.50	66.91
IAN (Ma et al., 2017)	72.10	-	78.60	-	-	-
RAM (Chen et al., 2017)	75.01	70.51	79.79	68.86	71.88	70.33
BILSTM-ATT-G	73.12	69.80	79.73	69.25	70.38	68.37
MGAN (Fan et al., 2018)	75.39	72.47	81.25	71.94	72.54	70.81
TNet (Li et al., 2018)	76.54	71.75	80.79	71.27	74.97	73.60
PBAN (Gu et al., 2018)	74.12	-	81.16	-	-	-
Cabasc (Liu et al., 2018)	75.07	-	80.89	-	71.53	-
IACapsNet	**76.80**	**73.29**	**81.79**	**73.40**	**75.01**	**73.81**

Table 1: Experimental results. The baseline results are retrieved from the original papers and (Li et al., 2018)

out-of-vocabulary words are initialized by sampling from the uniform distribution $U(-0.1, 0.1)$. In the Bi-RNN layer, LSTM is utilized and the dimension of the hidden state is 300. The dimension of the capsule's output is 16. We use Adam optimizer (Kingma and Ba, 2014) as our optimization method with $5e-4$ learning rate. L2 regularization and dropout are adopted to avoid overfitting.

Dataset	Positive		Neural		Negative	
	Train	Test	Train	Test	Train	Test
Restaurant	2164	728	637	196	807	196
Laptop	994	341	464	169	870	128
Twitter	1561	173	3127	346	1560	173

Table 2: Summary statistics of the datasets.

4.2 Model Comparisons

IACapsNet is compared with the following methods:

ATAE-LSTM(Wang et al., 2016a): An LSTM-based model which learns attention embeddings and combine them with the LSTM hidden states to predict the polarity.

TD-LSTM (Tang et al., 2016b): It employs two LSTMs to estimate the left context and the right context, respectively. The concatenated context representations perform the predictions.

IAN (Ma et al., 2017): An interactive attention is implemented on the representation of context and aspect learned by two LSTMs.

MemNet (Tang et al., 2016c): It applies attention mechanism over the word embeddings multiple times and predicts sentiment based on the topmost sentence representation.

RAM (Chen et al., 2017): Similar to MemNet, RAM is a multi-layer architecture where each layer consists of attention-based aggregation of word features and a GRU cell to learn the sentence representation.

BILSTM-ATT-G: (Zhang and Liu, 2017): It models left and right contexts using two attention-based LSTMs and introduces gates to measure the importance of left context, right context, and the entire sentence for the prediction.

MGAN (Fan et al., 2018): MGAN leverages the fine-grained and coarse-grained attention, which is further employed to characterize the word-level interactions between aspect and context words.

PBAN (Gu et al., 2018): PBAN concentrates on the position information of aspect terms and mutually models the relation between aspect term and sentence by employing bidirectional attention.

TNet (Li et al., 2018): It employs a CNN layer to extract salient features from the transformed word representations originated from a bi-directional RNN layer.

Cabasc (Liu et al., 2018): Cabasc employs sentence-level content attention mechanism to capture the important information about given aspects from a global perspective.

4.3 Main Results

As shown in Table 1, IACapsNet achieves the best performance on all the datasets. From Table 1, we can have the following observations.

ATAE-LSTM performs better than TD-LSTM.

One main reason may be the attention mechanism in TD-LSTM that enables to notice the important parts based on the aspect term. BILSTM-ATT-G adopts a similar architecture with TD-LSTM by modeling left context and right context using attention-based LSTM, which achieves better results than ATAE-LSTM. IAN and MGAN introduce the interactive attention in coarse-grained and multi-grained ways respectively and bring remarkable improvements. PBAN similarly utilize a fine-grained bidirectional attention and performs comparably with MGAN.

MemNet utilizes a more complex structure that contains nine computational layers, which updates the query vector at each hop. RAM also learns multiple attended vectors on the memory, which achieves superior results among the baseline models, especially on Laptop dataset.

Our proposed IACapsNet consistently performs best on all the three datasets. The improvement is mainly attributed to the feature clustering ability to tackle the overlapped features and the iteratively updating on coupling coefficients, which considers the overall meaning of the contexts. Moreover, compared with Cabasc, which also incorporates the overall representation to typical attention mechanism in a static way, our iterative method shows remarkable strengths.

4.4 Ablation Study

To analyze the effect of different components including the routing mechanism and the introduced interactive attention, we report the results of variants of IACapsNet. The results in Table 3 indicate: (1) EM routing based IACapsNet outperforms IACapsNet-Cosine, which routes capsules by cosine similarity (Sabour et al., 2017). One main reason maybe cosine saturates at 1, which is insensitive to the difference between a quite good agreement and a very good agreement. (2) Integrating interactive attention in routing mechanism brings a remarkable improvement on both routing mechanisms, which demonstrates the necessity to consider the relationship between the aspect and contexts during the routing procedure.

Moreover, EM routing also brings a boost on efficiency with fewer trainable parameters and faster speed which is intuitively shown in Table 4 (IAN is listed as baseline). From the table, it is easy to conclude that IACapsNet is much more efficient than IACapsNet-Cosine with about 10% and 36% decrease in the number of trainable parameters and running speed, respectively. Moreover, compared to IAN, IACapsNet achieves a much better accuracy with fewer trainable parameters, meaning that capsule network is more efficient in feature encoding with fewer parameters. However, IACapsNet costs more time compared to IAN, which is the implicit deficiency of capsule network because of the iterative calculation during routing.

4.5 Effects of Routing Iteration Number

As our proposed IACapsNet involves iterative procedure during routing. In this section, we investigate the effects of different routing iteration numbers. Specifically, we conduct experiments on all the three datasets and vary routing iteration numbers r from 1 to 4. The results are illustrated in Figure 3.

The results show that IACapsNet achieves the best performance at routing iteration number 2, 3 and 3 on the dataset RESTAURANT, LAPTOP, and Twitter, respectively. When the number of iteration is 1, our capsule network degrades to a standard network, which obtains comparable results with IAN. While increasing r to 4, the performance gets worse dramatically. Moreover, as the number of iteration increases to 4, it brings many difficulties to train IACapsNet. The model becomes more sensitive, which fluctuates greatly in loss and accuracy during training. Therefore, it is appropriate to limit the routing iteration number r and set it to be 2 or 3 depending on the performance.

Model	Accuracy	Macro-F1
IACapsNet-cosine routing	81.12	72.05
-w/o interactive attention	80.64	71.76
IACapsNet-EM routing	**81.79**	**73.40**
-w/o interactive attention	80.89	71.84

Table 3: Ablation study on Restaurant dataset. Cosine routing denotes routing measured by cosine similarity (Sabour et al., 2017), and w/o means without.

4.6 Case Study

In order to assess the effect of our EM routing with interactive attention mechanism, we visualize the coupling coefficients. Our model is able to adjust the contribution of each part based on the global meaning of a sentence and shows superiority in modeling complicated sentence. In this section, we pick an example from RESTAURANT dataset

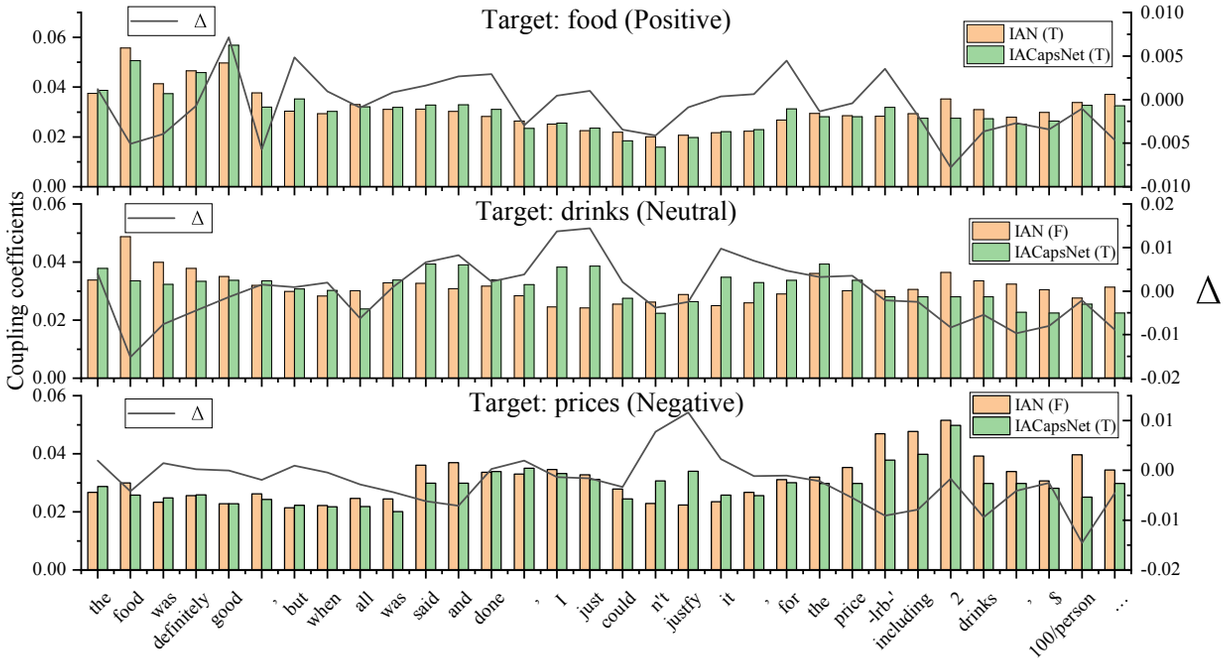

Figure 2: Routing Visualization

Model	Para.	Train	Infer.
IAN (Ma et al., 2017)	5.057	**0.11**	**0.017**
IACapsNet-cosine routing	4.614	0.33	0.044
IACapsNet	**4.189**	0.21	0.024

Table 4: Evaluation of efficiency. "Para." denotes the number of trainable parameters (M). "Train" and "Infer." denotes the training speed and inference speed(second/step). Speed is measured on 1 NVIDIA p100 GPU with batch size 64.

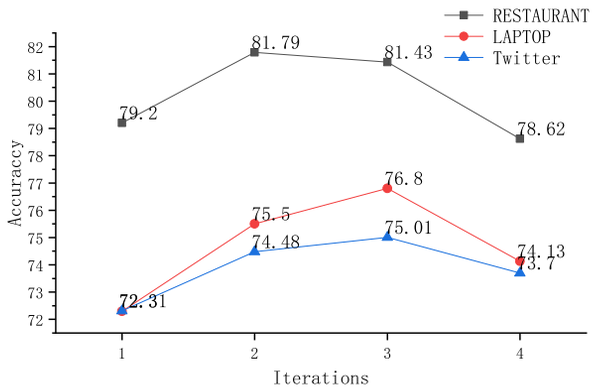

Figure 3: Effects of Routing Iteration Number

consisting of a long and complicated sentence and 3 aspect terms with 3 different sentimental polarities. Figure 2 shows this example and the visualization results of word-level coupling coefficients which are the sum of one word's coupling coefficients to the category. The attention weights from IAN are also shown as a baseline, which is normalized to the same scale with routing coupling coefficients. The line in the chart reflects the difference Δ. 'F' in Figure 2 means false sentiment classification, and 'T' means correct classification.

From the figure, we can observe that our routing methods can adjust the attended words according to different aspect terms, which helps make all the predictions correctly. Moreover, our routing method can locate on the more important words more efficiently. For example, in terms of the aspect 'price', the words "n't" and "justify" is attended, which are the corresponding essential sentimental words. However, they are ignored by the ordinary attention mechanism, which leads a wrong prediction. This shows that our routing mechanism can capture important parts of a sentence more accurately.

4.7 Conclusion and Future Work

We re-examine the deficiencies of existing models with attention mechanism for aspect-level sentiment classification. And we propose to utilize capsule network to handle the overlapped sentiment features by features clustering, and iteratively adjust the attention weights from a global perspective. To the best of our knowledge, capsule network is firstly applied in this task. Moreover, interactive attention is introduced to the dynamic rout-

ing to model the semantic relationship between aspect term and sentence. The experimental results verify that IACapsNet outperforms baseline models. The ablation and case studies show the efficacy of different proposed modules.

In the future, our theory can be generalized to other tasks that highly depends on the attention mechanism. For example, reading comprehension and machine translating.

Acknowledgements

This work was jointly supported by: (1) National Natural Science Foundation of China (No. 61771068, 61671079, 61471063, 61372120, 61421061); (2) Beijing Municipal Natural Science Foundation (No.4182041, 4152039); (3) the National Basic Research Program of China (No. 2013CB329102); (4) Fundamental Research Funds for the Central Universities under Grant 2018RC20; (5) BUPT Excellent Ph.D. Students Foundation.

References

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*.

Shuqin Gu, Lipeng Zhang, Yuexian Hou, and Yin Song. 2018. A position-aware bidirectional attention network for aspect-level sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*.

Geoffrey Hinton, Sara Sabour, and Nicholas Frosst. 2018. Matrix capsules with em routing. In *ICLR 2018*.

Geoffrey E. Hinton, Alex Krizhevsky, and Sida D. Wang. 2011. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning - ICANN 2011 - 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I*.

Binxuan Huang and Kathleen M. Carley. 2018. Parameterized convolutional neural networks for aspect level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR*.

Cheng Li, Xiaoxiao Guo, and Qiaozhu Mei. 2017. Deep memory networks for attitude identification. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, WSDM 2017, Cambridge, United Kingdom, February 6-10, 2017*.

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Qiao Liu, Haibin Zhang, Yifu Zeng, Ziqi Huang, and Zufeng Wu. 2018. Content attention model for aspect based sentiment analysis. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*.

Verónica Pérez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016b. Effective lstms for target-dependent sentiment classification. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*.

Duyu Tang, Bing Qin, and Ting Liu. 2016c. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*.

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

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

Wei Xue and Tao Li. 2018. Aspect based sentiment analysis with gated convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*.

Min Yang, Wei Zhao, Jianbo Ye, Zeyang Lei, Zhou Zhao, and Soufei Zhang. 2018. Investigating capsule networks with dynamic routing for text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*.

Ningyu Zhang, Shumin Deng, Zhanling Sun, Xi Chen, Wei Zhang, and Huajun Chen. 2018. Attention-based capsule network with dynamic routing for relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*.

Xinsong Zhang, Pengshuai Li, Weijia Jia, and Hai Zhao. 2019. Multi-labeled relation extraction with attentive capsule network. *AAAI*.

Yue Zhang and Jiangming Liu. 2017. Attention modeling for targeted sentiment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*.

Shenjian Zhao and Zhihua Zhang. 2018. Attention-via-attention neural machine translation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*.