Interactive Attention Transfer Network for Cross-domain Sentiment Classification

Kai Zhang†, Hefu Zhang†, Qi Liu‡§*, Hongke Zhao†, Hengshu Zhu‡, Enhong Chen†§

†Anhui Province Key Laboratory of Big Data Analysis and Application, University of Science and Technology of China
‡School of Data Science, University of Science and Technology of China
§Baidu Talent Intelligence Center, Baidu Inc

{sa517494, zhf2011, zhhk}@mail.ustc.edu.cn, {qiliuql, cheneh}@ustc.edu.cn, zhuhengshu@gmail.com

Kai Zhang 2018/1/28
Outline

1. Background
2. Related Work
3. IATN Framework
4. Experiment
5. Conclusion
01 Background
Background

• What is **Sentiment classification**?
  • Opinion mining
  • Opinion extraction
  • Sentiment analysis
  • Sentiment mining

Use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment from typically unstructured text.
Background

• **What is Transfer Learning?**
  • Refers to the use of similarities between tasks to apply the knowledge learned in the old field to a new field.

**Advantages:**
• Massive unlabeled data
• Cold start
• Model versatility
• Weak computing...

**Domain:**
• Source: with labeled data
• Target: less or no labeled data
Background

- **Cross-domain Sentiment Classification**
  - Refers to the use of similarities between text classification tasks to apply the knowledge learned in the source domain to the target domain.

**Problem**: Unsupervised Sentiment Classification

- From: labeled source domain data
- To: sentiment label (target domain)
- Focus on short-text mining and transfer learning methods.
02 Related Work
Related Work

• Past researches can be categorized into two perspectives.
  • Traditional (Theoretical) [Blitzer 2006; Pan et al. 2010; Chen et al. 2012)]
    • Analyzing the major features across domains.
  • Deep-learning Networks [Ganin et al. 2016, Li et al. 2017; 2018b]
    • Modeling the common sentiment relations between different domains.

However, most of the previous efforts ignore the characteristics which do not express the sentiment directly, such as the individual modeling of the aspect.
Related Work

- Passed researches can be categorized as **two perspectives**.
  - **Traditional (Theoretical)** [Blitzer 2006; Pan et al. 2010; Chen et al. 2012]
    - Analyzing the major features across domains.
  - **Deep-learning Networks** [Ganin et al. 2016, Li et al. 2017; 2018b]
    - Modeling the common relations between different domains.

However, most of the previous efforts ignore the characteristics which do not express the sentiment directly, such as the individual modeling of the aspect.

For Example: “the **voice quality** of iPhone is great, but its **battery** sucks”

- Domain shared-feature (pivots)
- Memory Network
- Hierarchical Attention Network
03 IATN Framework
IATN Framework

• Challenge One
  • What is the key information in the sentence that we deserve?

• Challenge Two
  • How do we make the most of this critical information?

The appearance of the PC looks good, but the battery life is too short.

The appearance of this dress looks nice, and the fabric is not bad.
IATN Framework

- **Aspect Extraction** [Li, Xin, et al 2018]

- **Method to extract aspect information?**
  - Truncated History Attention
  - Integrate the information of aspect detection history
  - Selective Transformation Network
  - Use opinion summary

---

Input sentences

1. the device speaks about it self
2. Great service !
3. Apple is unmatched in product quality, aesthetics, craftsmanship, and customer service
4. I am pleased with the fast log on, speedy WiFi connection and the long battery life
5. Also, I personally wasn’t a fan of the portobello and asparagus mole

Output of Ours

NONE

service

product quality, aesthetics, craftsmanship, customer service

log on, WiFi connection, battery life

portobello and asparagus mole
IATN Framework

• Problem Definition

Given two domains, $D_s$ is the source domain and $D_t$ is the target domain. Denote that each item (e.g., review) at both domains consists of n words marked as $s=\{w_s^1, w_s^2, ..., w_s^n\}$ and their aspect sequence contains m words marked as $a=\{w_a^1, w_a^2, ..., w_a^m\}$.

The goal is to train a robust model based on labeled data in source domain and adapt it to predict the unlabeled data in target domain.
IATN Framework

- Interactive Attention Transfer Network
  - Construct a better sentiment transfer network
  - Make full use of the information of the aspect

- Consist of two similar parts
  - S-net
    - Learn shared information between domains
    - Pooling feature vectors to A-net
  - A-net
    - Modeling aspect information
    - Use aspect information as the auxiliary task of domain classification.
IATN Framework

- **IATN – Embedding Layer**
  - After aspect extraction, we map each word into a low-dimensional real-value vector.
  - We choose the pre-training method and take each word to a embedding vector.

- **IATN – LSTM Layer**
  - We adopt LSTM to learn hidden states.
  - Performs well in learning long-term dependencies.
  - Solve gradient vanishing and expansion problems
IATN Framework

- **IATN – Embedding Layer**
  - After aspect extraction, we map each word into a low-dimensional real-value vector.
  - We choose the pre-training method and take each word to a embedding vector.

- **IATN – LSTM Layer**

\[
\begin{align*}
    i_t & = \delta(W_ei^t e_s + W_i h_{t-1} + \hat{b}_i), \\
    f_t & = \delta(W_eff e_s + W_f h_{t-1} + \hat{b}_f), \\
    c_t & = f_t \cdot c_{t-1} + i_t \cdot \tau(W_ec e_s + W_c h_{t-1} + \hat{b}_c), \\
    o_t & = \delta(W_eo e_s + W_o h_{t-1} + \hat{b}_o), \\
    h_t & = o_t \cdot \tanh(c_t),
\end{align*}
\]
IATN Framework

- **IATN – Pooling Layer**
  
  After getting the hidden state representations of sentences and aspects, we need to make them interactive.
  
  \[ h^p_s = \sum_{i=1}^{n} (h^i_s / n) \quad h^p_a = \sum_{i=1}^{m} (h^i_a / m) \]

- **IATN – Attention Layer**
  
  \[ \alpha_i = \frac{\exp(\gamma_s(h^i_s, h^p_a))}{\sum_{j=1}^{n+1} \exp(\gamma_s(h^i_s, h^p_a))} \]
  \[ \gamma_s(h^i_s, h^p_a) = \tanh([h^i_s, h^p_a] \cdot W_s + b_s) \]
  
  \[ S_r = \sum_{i=1}^{n} \alpha_i h^i_s \quad A_r = \sum_{i=1}^{n} \beta_i h^i_a \]

The same way from A-net
IATN Framework

- IATN – Two tasks
  - Domain Classification
    
    GRL-Layer: \( G_i = G(S_r) \).

    \[
    y_d' = \text{softmax}(G_i)
    \]

    \[
    Loss_{dom} = -\frac{1}{N} \sum_{i=1}^{N} y_d' \ln y_d + (1 - y_d') \ln(1 - y_d).
    \]

  - Sentiment Classification

    \[
    V_i = S_r \oplus A_r
    \]

    \[
    y_s' = \text{softmax}(V_i)
    \]

    \[
    Loss_{sen} = -\frac{1}{N} \sum_{i=1}^{N} y_s' \ln y_s + (1 - y_s') \ln(1 - y_s).
    \]

    Minimize: \( L = L_{sen} + L_{dom} + \rho L_{reg} \)
04 Experiment
• Data Description
  • Amazon dataset (Main experiment)
  • Indiegogo dataset (Application verification)

(Download) http://www.indiegogo.com/download/

Table 1: Statistics of datasets after pre-processing.

<table>
<thead>
<tr>
<th>Domains</th>
<th># Train</th>
<th># Test</th>
<th># Unlabel</th>
</tr>
</thead>
<tbody>
<tr>
<td>Books</td>
<td>5,000</td>
<td>1,000</td>
<td>8,000</td>
</tr>
<tr>
<td>DVD</td>
<td>5,000</td>
<td>1,000</td>
<td>8,000</td>
</tr>
<tr>
<td>Electronics</td>
<td>5,000</td>
<td>1,000</td>
<td>8,000</td>
</tr>
<tr>
<td>Kitchen</td>
<td>5,000</td>
<td>1,000</td>
<td>8,000</td>
</tr>
</tbody>
</table>

• Amazon data preparation:
  we choose the Amazon reviews data from four domains: Book, DVD, Electronics and Kitchen appliances. Each of the domains contains 6,000 labeled data.
**Experiment**

- **Data Analysis**
  - We observe that similar domains have more same aspects than different domains.
    (Which can prove something latter!)
  - B and D domain have shared most amount of aspects (plots, character, story, ...)
  - B and K domain shared the least amount of aspects (those two domain are too far apart)

Top-100 aspects analysis between domains:
- We selected the top 100 Aspects with the highest frequency between every two domains.
**Experiment**

- **Benchmarks**
  - **Naive** is a non-domain-adaptive method which is trained in the source domain and predicts in target domain directly. It is designed based on LSTM.
  - **SCL** is a linear method, which aims to solve feature mismatch problem by aligning domain common and unique features.
  - **SFA** is a method which aims to build a bridge between the source and the target domains by aligning common and unique features.
  - **mSDA** is proposed to automatically learn a unified feature representation for sentences from a large amount of data in all the domains.
  - **DANN** is based on the adversarial training. DANN performs domain adaptation with the representation encoded in a 5000-dimension feature vector.
  - **CNN-aux** is based on Convolutional Neural Network and makes use of two auxiliary tasks to help inducing sentence embedding.
  - **AMN** is a method which learns domain-shared representations based on memory networks and adversarial training.
  - **HATN & HATN<sup>h</sup>** are hierarchical attention networks to focus on both the word level and the sentence level sentiment. The former one does not contain the hierarchical positional encoding and the latter one does.
Experiment

• Results

• As the table shows, IATN model has achieved the best performances on most tasks of this dataset.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>B→D</th>
<th>B→E</th>
<th>B→K</th>
<th>D→B</th>
<th>D→E</th>
<th>D→K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.786</td>
<td>0.752</td>
<td>0.737</td>
<td>0.756</td>
<td>0.734</td>
<td>0.767</td>
</tr>
<tr>
<td>SCL</td>
<td>0.807</td>
<td>0.763</td>
<td>0.771</td>
<td>0.782</td>
<td>0.754</td>
<td>0.779</td>
</tr>
<tr>
<td>SFA</td>
<td>0.813</td>
<td>0.776</td>
<td>0.785</td>
<td>0.788</td>
<td>0.758</td>
<td>0.786</td>
</tr>
<tr>
<td>mSDA</td>
<td>0.819</td>
<td>0.783</td>
<td>0.789</td>
<td>0.783</td>
<td>0.770</td>
<td>0.793</td>
</tr>
<tr>
<td>DANN</td>
<td>0.832</td>
<td>0.764</td>
<td>0.790</td>
<td>0.805</td>
<td>0.796</td>
<td>0.814</td>
</tr>
<tr>
<td>CNN-a</td>
<td>0.843</td>
<td>0.810</td>
<td>0.813</td>
<td>0.829</td>
<td>0.803</td>
<td>0.819</td>
</tr>
<tr>
<td>AMN</td>
<td>0.855</td>
<td>0.824</td>
<td>0.811</td>
<td>0.846</td>
<td>0.812</td>
<td>0.827</td>
</tr>
<tr>
<td>HATN</td>
<td>0.858</td>
<td>0.853</td>
<td>0.849</td>
<td>0.858</td>
<td>0.849</td>
<td>0.853</td>
</tr>
<tr>
<td>HATN&lt;sup&gt;h&lt;/sup&gt;</td>
<td>0.861</td>
<td>0.857</td>
<td>0.852</td>
<td>0.863</td>
<td>0.856</td>
<td>0.862</td>
</tr>
<tr>
<td>IATN&lt;sup&gt;n&lt;/sup&gt;</td>
<td>0.854</td>
<td>0.849</td>
<td>0.838</td>
<td>0.848</td>
<td>0.855</td>
<td>0.839</td>
</tr>
<tr>
<td>IATN</td>
<td>0.868</td>
<td>0.865</td>
<td>0.859</td>
<td>0.870</td>
<td>0.869</td>
<td>0.858</td>
</tr>
</tbody>
</table>

From the experimental results, we can also observe that the classification accuracy between similar domains will be higher than different domain. For example, “B&D” task is more accurate than “B&K” task because they have more similar aspects. (As we mentioned before!)
Experiment

Case Study

In order to validate that our model is able to identify the impact of aspect on sentiment representation, we visualize the aspect attention layer.

We can see that although “acting” and “pace” both appear with positive emotions in this review, but the more effective aspects “character” and “plot” have received negative emotions. Affected by these more powerful aspects, this review is finally labeled as negative. (Aspect influence different across domains!)

Book-pos Example:
This story captivated me right from the outset, as Vivian vague of memory. Her scenes are drawn in sensitively described and insightful detail, and she is a very realistically portrayed.

Book-neg Example:
I am not sure whatever possessed me to buy this book. Honestly, it was a complete waste of free time. To quote a friend, it was not the best use of my entertainment dollar. If you are a fan of pedestrian writing, lack-luster plots and hackneyed character development, this is your book.

Dvd-pos Example:
An amazing film! When seeing this movie’s poster, i was not too excited, but when watching it realized how awesome it really is. Not only it’s story is well laid out, but the amount of special effects, great scenes and good actors.

Dvd-neg Example:
The acting was good. The pace was adequate. However, the plot was predictable. The movie content just reeks of the intelligent thriller syndrome. Clive’s character kept calling this the perfect robbery. I’m still don’t understand what Jody Foster’s character brings to the plot.

Aspect colored in deeper green means that it gains the heavier weight.
Experiment

• Application Verification

Prediction Model
• The further mission of cross-domain sentiment classification aims to solve the problem of unlabeled domain.
• Thus, to further verify the effectiveness of IATN, we design an application on an unlabeled dataset. (Crowdfunding review dataset!)

Prediction Result
• As Table shows, the SVM method without review information performs the worst and the accuracy result is 5.11% lower than LSTM\textsuperscript{iatn}.
• LSTM\textsuperscript{iatn} improves the accuracy by 3.20% and 2.42% than LSTM\textsuperscript{one} and LSTM\textsuperscript{hatn} respectively.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.7671</td>
<td>0.4567</td>
<td>0.6971</td>
<td>0.5536</td>
</tr>
<tr>
<td>LSTM\textsuperscript{one}</td>
<td>0.7862</td>
<td>0.4813</td>
<td>0.6732</td>
<td>0.5689</td>
</tr>
<tr>
<td>LSTM\textsuperscript{hatn}</td>
<td>0.7940</td>
<td>0.4843</td>
<td>0.6805</td>
<td>0.5674</td>
</tr>
<tr>
<td>LSTM\textsuperscript{iatn}</td>
<td><strong>0.8182</strong></td>
<td><strong>0.4977</strong></td>
<td><strong>0.6733</strong></td>
<td><strong>0.5743</strong></td>
</tr>
</tbody>
</table>
05 Conclusion
Conclusion

• What we do?
  • We propose to make cross-domain sentiment classification by integrating the contextual representation and aspect representation of sentences.
  • We propose a novel IATN method which utilizes the interactive attention mechanism to get important information from both the sentence and aspect.
  • We conduct extensive experiments on two real-world datasets. The experimental results validate that our method outperforms other state-of-the-art methods.

• Future
  • Add aspect information to the hierarchical attention network.
  • Do some sentiment transfer learning at the aspect-level.
Thank you!  Q&A

Kai Zhang
2018/1/28