Efficient Adaptive Transfer Network for Aspect-level Sentiment Analysis

Kai Zhang    2021/04/08
01 Background
Background

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

- **Sentiment classification** is a special task of text classification
- It’s objective is to **classify** a text according to the sentimental polarities of opinions it contains (Pang et al., 2002)
- E.g., favorable or unfavorable, positive or negative.

Emotion Tendency

Attitude?

- Words
- Pictures
- Behaviors
Background

• What is **Aspect-level Sentiment classification**?
  • A more granular sentiment classification task which refers to identifying sentiment polarities towards aspects in a sentence.
  • E.g. (Aspect\(^1\) words)
    • Color; battery life
    • Service; food

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

The **service**, as a matter of fact, is not as good as I think, but I still like their **food**.

\(^1\) aspect is also regards as target.
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

• **Why Transfer Learning?**
  - Myth: Unless your question has one million labeled data, you cannot do deep learning.
  - Save costs and improve efficiency!

![Transfer Learning](https://www.youtube.com/watch?v=wjqaz6m42wU&feature=youtu.be)
02 Aspect-level Sentiment Classification
Aspect-level Sentiment Classification

• Effective LSTMs for Target-Dependent Sentiment Classification
  • Tang D et al. (COLING 2016) proposed TD-LSTM, TC-LSTM

  \[ \text{LSTM}_L : \text{the preceding contexts plus target}^{1} \text{ string (left to right)} \]
  \[ \text{LSTM}_R : \text{target string plus the following contexts (right to left)} \]
  • regarding target string as the last unit could better utilize the semantics of target string

\(^1\) aspect is also regards as target.
Aspect-level Sentiment Classification

- Effective LSTMs for Target-Dependent Sentiment Classification
  - Tang D et al. (COLING 2016) proposed TD-LSTM, TC-LSTM

- Target-dependent LSTM (TD-LSTM) could make better use of the target information
- Not good enough to capture the interactions between target word and its contexts
Aspect-level Sentiment Classification

• Effective LSTMs for Target-Dependent Sentiment Classification

• Experiment

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM-indep</td>
<td>0.627</td>
<td>0.602</td>
</tr>
<tr>
<td>SVM-dep</td>
<td>0.634</td>
<td>0.633</td>
</tr>
<tr>
<td>Recursive NN</td>
<td>0.630</td>
<td>0.628</td>
</tr>
<tr>
<td>AdaRNN-w/oE</td>
<td>0.649</td>
<td>0.644</td>
</tr>
<tr>
<td>AdaRNN-w/E</td>
<td>0.658</td>
<td>0.655</td>
</tr>
<tr>
<td>AdaRNN-comb</td>
<td>0.663</td>
<td>0.659</td>
</tr>
<tr>
<td>Target-dep</td>
<td>0.697</td>
<td>0.680</td>
</tr>
<tr>
<td>Target-dep+</td>
<td>0.711</td>
<td>0.699</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.665</td>
<td>0.647</td>
</tr>
<tr>
<td>TD-LSTM</td>
<td>0.708</td>
<td>0.690</td>
</tr>
<tr>
<td>TC-LSTM</td>
<td><strong>0.715</strong></td>
<td><strong>0.695</strong></td>
</tr>
</tbody>
</table>

Example

<table>
<thead>
<tr>
<th>Example</th>
<th>gold</th>
<th>LSTM</th>
</tr>
</thead>
<tbody>
<tr>
<td>i hate my ipod look at my last tweet before the argh one that ’s for you</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>okay soooo ... ummmmm .... what is going on with lindsay lohan’ s face?</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>boring day at the office = perez and tomorrow overload. not good</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>i heard ShannonBrown did his thing in the lakers game!! got ta love him</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Hey google, thanks for all these great Labs features on Chromium, but how about ” Create Application Shortcut”?!</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Best performance!

whole sentence
Aspect-level Sentiment Classification

- Relational Graph Attention Network for Aspect-based Sentiment Analysis
  - Wang K et al. (ACL 2020) proposed R-GAT
    - Not all words contribute the same point
    - The purpose is to better model neighbor words

\[
h_{att}^{l+1} = \|_{k=1}^{K} \sum_{j \in \mathcal{N}_i} \alpha_{ij}^{lk} W_k^l h_j^l
\]

\[
\alpha_{ij}^{lk} = \text{attention}(i, j)
\]

- GAT:

- R-GAT:

\[
h_{rel}^{l+1} = \|_{m=1}^{M} \sum_{j \in \mathcal{N}_i} \beta_{ij}^{lm} W_m^l h_j^l
\]

\[
g_{ij}^{lm} = \sigma(\text{relu}(r_{ij} W_{m1} + b_{m1}) W_{m2} + b_{m2})
\]

\[
\beta_{ij}^{lm} = \frac{\exp(g_{ij}^{lm})}{\sum_{j=1}^{N_i} \exp(g_{ij}^{lm})}
\]
Aspect-level Sentiment Classification

- Relational Graph Attention Network for Aspect-based Sentiment Analysis

  ➢ Concat:

  \[ x_i^{l+1} = h_{att_i}^{l+1} \parallel h_{rel_i}^{l+1} \]

  \[ h_i^{l+1} = \text{relu}(W_{l+1}x_i^{l+1} + b_{l+1}) \]

  ➢ Prediction:

  \[ p(a) = \text{softmax}(W_p h_a^l + b_p) \]

- Dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Positive</th>
<th>Neutral</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Train</td>
<td>Test</td>
<td>Train</td>
</tr>
<tr>
<td>Laptop</td>
<td>994</td>
<td>341</td>
<td>870</td>
</tr>
<tr>
<td>Restaurant</td>
<td>2164</td>
<td>728</td>
<td>807</td>
</tr>
<tr>
<td>Twitter</td>
<td>1561</td>
<td>173</td>
<td>3127</td>
</tr>
</tbody>
</table>
### Aspect-level Sentiment Classification

- **Relational Graph Attention Network for Aspect-based Sentiment Analysis**

- **Results**

<table>
<thead>
<tr>
<th>Category</th>
<th>Method</th>
<th>Restaurant</th>
<th>Laptop</th>
<th>Twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
</tr>
<tr>
<td>Syn.</td>
<td>LSTM + SynATT</td>
<td>80.45</td>
<td>71.26</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>AdaRNN</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>PhraseRNN</td>
<td>66.20</td>
<td>59.32</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ASGCN</td>
<td>80.77</td>
<td>72.02</td>
<td>75.55</td>
</tr>
<tr>
<td></td>
<td>CDT</td>
<td>82.30</td>
<td>74.02</td>
<td>77.19</td>
</tr>
<tr>
<td></td>
<td>GAT</td>
<td>78.21</td>
<td>67.17</td>
<td>73.04</td>
</tr>
<tr>
<td></td>
<td>TD-GAT</td>
<td>80.35</td>
<td>76.13</td>
<td>74.13</td>
</tr>
<tr>
<td>Att.</td>
<td>EATAE-LSTM</td>
<td>77.20</td>
<td>-</td>
<td>68.70</td>
</tr>
<tr>
<td></td>
<td>EAN</td>
<td>78.60</td>
<td>-</td>
<td>72.10</td>
</tr>
<tr>
<td></td>
<td>EARM</td>
<td>80.23</td>
<td>70.80</td>
<td>74.49</td>
</tr>
<tr>
<td></td>
<td>EMGAN</td>
<td>81.25</td>
<td>71.94</td>
<td>75.39</td>
</tr>
<tr>
<td></td>
<td>ESTM</td>
<td>79.10</td>
<td>69.00</td>
<td>71.22</td>
</tr>
<tr>
<td></td>
<td>BERT</td>
<td><strong>85.62</strong></td>
<td><strong>78.28</strong></td>
<td><strong>77.58</strong></td>
</tr>
<tr>
<td>Others</td>
<td>GCAE</td>
<td>77.28</td>
<td>-</td>
<td>69.14</td>
</tr>
<tr>
<td></td>
<td>JCI</td>
<td>-</td>
<td>68.84</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>VNET</td>
<td>80.69</td>
<td>71.27</td>
<td>76.54</td>
</tr>
<tr>
<td>Ours</td>
<td>R-GAT</td>
<td>83.30</td>
<td>76.08</td>
<td>77.42</td>
</tr>
<tr>
<td>Ours</td>
<td>R-GAT + BERT</td>
<td><strong>86.60</strong></td>
<td><strong>81.35</strong></td>
<td><strong>78.21</strong></td>
</tr>
</tbody>
</table>
Limitations of Previous Works:

- Existing aspect-level sentiment classification approaches generally apply a uniform model across domains.
- Due to the explosive growth of the online social media data, there exists a barrier for applying the aspect-level sentiment classifier to a new domain which has few labeled data.
- Labeling large-scale data may be time-consuming and expensive.
03 Aspect-level Sentiment Classification — Considering Transferability
Efficient Adaptive Transfer Network

• Challenges

• How to transfer knowledge across domains?
• How to learn aspect-aware semantic information during domain transfer?
• How to reduce the domain discrepancy during domain transfer?
Multi-head Attention Transfer Network

- Problem Definition

Given two domains, $\mathcal{D}^s = \{x_i^s, a_i^s, y_i^s\}_{i=1}^{n_s}$ is the source domain and $\mathcal{D}^t = \{x_j^t, a_j^t\}_{j=1}^{n_t}$ is the target domain. Denote that each item (e.g., review) at both domains consists of $n$ context words marked as $s = \{w_1^c, w_2^c, ..., w_n^c\}$ and their aspect sequence contains $m$ words marked as $a = \{w_1^a, w_2^a, ..., w_m^a\}$.

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.

Cross-domain sentiment classification

- Source
- Labeled
- Model
- Knowledge
- Target
- Unlabeled
- Model
- Prediction
EATN Framework

• An overview
  • We propose a deep adaptive model named Efficient Adaptive Transfer Network, incorporating with multi-module to solve the problems mentioned above.

• EATN mainly contains three components
  • Embedding Module
  • Aspect-oriented Multi-head Attention Module
  • Domain Adaptation Module (DAM)
Embedding Module

- **Input Preprocessing**
  - One item may include multiple aspects, we split the sentence into multiple ones which each consist of just one aspect.

- **Word Embedding**
  - bert, google-research
  - bert-as-service

\[
\{w_1^c, w_2^c, \ldots, w_n^c\} \quad \{w_1^a, w_2^a, \ldots, w_m^a\}
\]
\[
\{e_1^c, e_2^c, \ldots, e_n^c\} \quad \{e_1^a, e_2^a, \ldots, e_m^a\}
\]

\[
H = [e_c \oplus e_a] = [e_1^c, e_2^c, \ldots, e_n^c, e_1^a, e_2^a, \ldots, e_m^a]
\]

"The customer service, as a matter of fact, is not as good as I think, but I still like their food" 

"customer service"—"The $, as a matter of fact, is not as good as I think, but I still like their food"

"food"—"The customer service, as a matter of fact, is not as good as I think, but I still like their $".
Aspect-oriented Multi-head Attention Module

• Purpose

• Mining the direct relationship between aspects and context;
• The *customer service*, as a matter of fact, is *not as good* as I think, but I still like their food.
• “as, matter, of, fact” are unrelated words.

\[ Q, K, V = HW^Q, HW^K, HW^V, \]
\[ Z = Attention(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_q}} \right) \]
\[ Q_i, K_i, V_i = QW_i^Q, KW_i^K, VW_i^V, \]
\[ Z_i = Attention(Q_i, K_i, V_i), \]
\[ O = [Z_1 \oplus Z_2 \oplus ... \oplus Z_t]. \]
Domain Adaptation Module

• Purpose
  • MLP to learn some deep latent features (high-dimensional representations)
  • How EATN get transferability progressively through Domain Adaptation Module (DAM)

• Consists of three stages
  • MLP layer with an MK-MMD (Long et al. ICML’15)
  • sentiment classification
  • domain classification
Domain Adaptation Module

- MLP layer with an MK-MMD
  - Deep features eventually convert from general to specific along with the network and the transferability gap grows
  - making the feature distribution similarly under each hidden layer

\[
R_1 = \text{Relu}(W_1O + b_1), \\
R_2 = \text{Relu}(W_2R_1 + b_2), \\
\vdots \\
R_h = \text{Relu}(W_hR_{h-1} + b_h),
\]

\[
d_k^2(p, q) \overset{\Delta}{=} \| \mathbf{E}_p - \mathbf{E}_q \|_{\mathcal{H}_k}^2,
\]

\[
\mathcal{L}_m = \sum_{\ell=1} \, d_k^2(R^s_{\ell}, R^t_{\ell}),
\]
Domain Adaptation Module

- **Sentiment Classification**
  - Focus on mining aspect-aware semantic information in the contextual sentences
    \[
    \hat{y}^s = \text{softmax}(W^s R_n^s + b^s). \quad \mathcal{L}_s = -\frac{1}{n_s} \sum_{i=1}^{n_s} (y_i^s \ln \hat{y}_i^s + (1 - y_i^s) \ln(1 - \hat{y}_i^s))
    \]

- **Domain Classification**
  - This task simultaneously optimizes the domain invariance to learn domain-shared features and to facilitate knowledge transfer across domains
    \[
    \hat{y}^d = \text{softmax}(W^d R_h + b^d). \quad \mathcal{L}_d = -\frac{1}{N} \sum_{i=1}^{N} (y_i^d \ln \hat{y}_i^d + (1 - y_i^d) \ln(1 - \hat{y}_i^d))
    \]

- **Loss Function**
  - Joint learning for them to optimize the parameters
    \[
    \mathcal{L} = \mathcal{L}_s + \mathcal{L}_d + \beta \mathcal{L}_m + \rho \mathcal{L}_{reg}
    \]
04 Experiment
## Experiment

- **Data Description**
  - Amazon dataset
  - Twitter dataset

- **Amazon data preparation:**
  
The first two are from the SemEval 2014 Task 4 (Kiritchenko et al. 2014), which contain reviews from Restaurants (R) and Laptop (L), respectively.

(Download) [http://alt.qcri.org/semeval2014/task4](http://alt.qcri.org/semeval2014/task4)

<table>
<thead>
<tr>
<th>Domains</th>
<th>Testing Set Percentage (20%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Pos.</td>
</tr>
<tr>
<td>Restaurant</td>
<td>2,890</td>
</tr>
<tr>
<td>Laptop</td>
<td>1,320</td>
</tr>
<tr>
<td>Twitter</td>
<td>1,800</td>
</tr>
</tbody>
</table>

- we conduct the cross-domain experiments between every two datasets, which means that we have six domain adaptation tasks: R→L,..., T→L. For example, the notation “R→L” represents the task which transfers from the source domain “Restaurant” to the target domain “Laptop”.
## Experiment

### Results

<table>
<thead>
<tr>
<th>Benchmarks</th>
<th>Res. →</th>
<th>Lap. →</th>
<th>Twitter. →</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R→L</td>
<td>R→T</td>
<td>L→R</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.6227(-13.2%)</td>
<td>0.5659(-18.9%)</td>
<td>0.5626(-11.6%)</td>
</tr>
<tr>
<td>TD-LSTM</td>
<td>0.6301(-12.9%)</td>
<td>0.5687(-19.0%)</td>
<td>0.5769(-10.4%)</td>
</tr>
<tr>
<td>ATAE</td>
<td>0.6023(-15.3%)</td>
<td>0.5839(-17.1%)</td>
<td>0.5652(-12.2%)</td>
</tr>
<tr>
<td>IAN</td>
<td>0.6280(-14.1%)</td>
<td>0.5814(-18.7%)</td>
<td>0.6371(-8.0%)</td>
</tr>
<tr>
<td>MemNet</td>
<td>0.6361(-14.5%)</td>
<td>0.5829(-19.8%)</td>
<td>0.6217(-9.6%)</td>
</tr>
<tr>
<td>AOA</td>
<td>0.6529(-10.7%)</td>
<td>0.5642(-29.5%)</td>
<td>0.6508(-8.7%)</td>
</tr>
<tr>
<td>MGNet</td>
<td>0.6541(-11.1%)</td>
<td>0.5920(-17.3%)</td>
<td>0.6631(-8.3%)</td>
</tr>
<tr>
<td>TNet</td>
<td>0.6642(-12.0%)</td>
<td>0.6099(-17.4%)</td>
<td>0.6711(-8.3%)</td>
</tr>
<tr>
<td>BERT-PT</td>
<td>0.6791(-12.3%)</td>
<td>0.6183(-18.3%)</td>
<td>0.6820(-9.7%)</td>
</tr>
<tr>
<td>SFA</td>
<td>0.6547</td>
<td>0.5924</td>
<td>0.6693</td>
</tr>
<tr>
<td>HATN</td>
<td>0.6762</td>
<td>0.6121</td>
<td>0.6887</td>
</tr>
<tr>
<td>IATN</td>
<td>0.6833</td>
<td>0.6169</td>
<td>0.6953</td>
</tr>
<tr>
<td><strong>EATN</strong></td>
<td><strong>0.7087</strong></td>
<td><strong>0.6323</strong></td>
<td><strong>0.7162</strong></td>
</tr>
</tbody>
</table>
## Experiment

### Results

<table>
<thead>
<tr>
<th>Benchmarks</th>
<th>Beas. $\rightarrow$</th>
<th>Hotel. $\rightarrow$</th>
<th>Res1. $\rightarrow$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B$\rightarrow$H</td>
<td>B$\rightarrow$R1</td>
<td>R1$\rightarrow$B</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.6695(-5.65%)</td>
<td>0.6753(-5.07%)</td>
<td>0.6983(-2.12%)</td>
</tr>
<tr>
<td>TD-LSTM</td>
<td>0.6766(-4.73%)</td>
<td>0.6709(-5.30%)</td>
<td>0.7078(-1.46%)</td>
</tr>
<tr>
<td>ATAE</td>
<td>0.6795(-5.37%)</td>
<td>0.6776(-5.56%)</td>
<td>0.7083(-2.26%)</td>
</tr>
<tr>
<td>IAN</td>
<td>0.6805(-2.07%)</td>
<td>0.6626(-3.84%)</td>
<td>0.7112(-2.14%)</td>
</tr>
<tr>
<td>MemNet</td>
<td>0.6923(-4.03%)</td>
<td>0.6790(-5.35%)</td>
<td>0.7208(-1.38%)</td>
</tr>
<tr>
<td>AOA</td>
<td>0.6529(-10.7%)</td>
<td>0.4642(-29.5%)</td>
<td>0.7136(-6.30%)</td>
</tr>
<tr>
<td>MGNet</td>
<td>0.6907(-3.59%)</td>
<td>0.6903(-3.62%)</td>
<td>0.7216(-0.41%)</td>
</tr>
<tr>
<td>TNet</td>
<td>0.7007(-4.06%)</td>
<td>0.7110(-3.02%)</td>
<td>0.7291(-1.73%)</td>
</tr>
<tr>
<td>BERT-PT</td>
<td>0.7172(-3.85%)</td>
<td>0.7194(-3.61%)</td>
<td>0.7411(-1.03%)</td>
</tr>
<tr>
<td>SFA</td>
<td>0.6860</td>
<td>0.6722</td>
<td>0.6817</td>
</tr>
<tr>
<td>HATN</td>
<td>0.6911</td>
<td>0.7138</td>
<td>0.7174</td>
</tr>
<tr>
<td>IATN</td>
<td>0.7074</td>
<td>0.7209</td>
<td>0.7227</td>
</tr>
<tr>
<td><strong>EATN</strong></td>
<td><strong>0.7218</strong></td>
<td><strong>0.7346</strong></td>
<td><strong>0.7410</strong></td>
</tr>
</tbody>
</table>

Anhui Province Key Laboratory of Big Data Analysis and Application, USTC
## Experiment

- **Case study**

<table>
<thead>
<tr>
<th>Cases are shown as below</th>
<th>Models and Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Example One.</strong></td>
<td></td>
</tr>
<tr>
<td>Aspects: “the cord” (Ground-truth: Neutral)</td>
<td>BERT-PT</td>
</tr>
<tr>
<td>Sentence: I charge it at night and skip taking “$” with me because of the good battery life.</td>
<td>Pos.</td>
</tr>
<tr>
<td><strong>Example Two.</strong></td>
<td></td>
</tr>
<tr>
<td>Aspects: “patches” (Ground-truth: Negative)</td>
<td></td>
</tr>
<tr>
<td>Sentence: 2nd Best computer in the world only one way this computer might become the best is that it needs to upgrade “$” to make less easier for people to hack into.</td>
<td>Neg.</td>
</tr>
<tr>
<td><strong>Example Three.</strong></td>
<td></td>
</tr>
<tr>
<td>Aspects: “meal” (Ground-truth: Positive)</td>
<td></td>
</tr>
<tr>
<td>Sentence: I just wonder how you can have such a delicious “$” for such little money.</td>
<td>Neg.</td>
</tr>
</tbody>
</table>
Future Work

• What we do?
  • We want to do cross-domain sentiment classification in aspect-level
  • We propose a novel EATN method which utilizes the multi-head attention mechanism to get direct relationship from aspect and it’s sentiment words
  • We design a Domain Adaptation Module to reduce the domain discrepancy. The experimental results validate that our method outperforms other state-of-the-art methods

• Future
  • Add transfer information to the hierarchical attention network
  • Analysis where to transfer and how to transfer.
05 Conclusion
References

Thank you!  Q&A

Kai Zhang
2021/04/08