X-Class: Text Classification with Extremely Weak Supervision

NAACL2021

Zihan Wang, Dheeraj Mekala, Jingbo Shang

UCSD
Overview

● Task
  ○ Text classification with extremely weak supervision, i.e., only relying on the surface text of class names.

● Key insights
  ○ Ideal document representations should lead to very close results between clustering and the desired classification
    ■ i.e., doc embeddings should reflect class info in clustering

Figure 1: Visualizations of News using Average BERT Representations. Colors denote different classes.
Figure 2: An overview of our X-Class. Given a raw input corpus and user-specified class names, we first estimate a class-oriented representation for each document. And then, we align documents to classes with confidence scores by clustering. Finally, we train a supervised model (e.g., BERT) on the confident document-class pairs.
M1: Class-oriented Document Representation

- **Class Representation Estimation**
  - Weighted average representation based on a ranked list of keywords
  - Incrementally add new keywords to list by ranking similarities of out-of-list words
    - Top-ranked keywords are expected to have more similar static representations to the class representation
  - Stop condition
    - New class rep. Changed the current list OR reach max T

- **Document Representation Estimation**
  - 4 ways to compute attention weight
    - 2 token rep. + 2 attention mechanisms
  - A unified list of geometric mean of the 4 ranks
    - Assign a weight of $1/r$ to a token ranked at $r$-th position

![Figure 3: Overview of Our Class Rep. Estimation.](image1)

![Figure 4: Overview of Our Document Rep. Estimation.](image2)
M1: Class-oriented Document Representation

- **Class Representation Estimation**
  - Weighted average representation based on a ranked list of keywords
  - Incrementally add new keywords to list by ranking similarities of out-of-list words
    - Top-ranked keywords are expected to have more similar static representations to the class representation
  - Stop condition
    - New class rep. Changed the current list OR reach max T

- **Document Representation Estimation**
  - 4 ways to compute attention weight
    - 2 token rep. + 2 attention mechanisms
  - A unified list of geometric mean of the 4 ranks
    - Assign a weight of $1/r$ to a token ranked at $r$-th position

---

**Algorithm 1: Class-Oriented Document Representation Estimation**

```
Input: n documents $D_i$, k class names $c_j$, max number of iterations $T$, and attention mechanism set $\mathcal{M}$
Output: Document representations $E_i$
Compute $t_{i,j}$ (contextualized token rep.)
Compute $s_w$ for all words (Eq. 1)

// class rep. estimation
for $j = 1 \ldots k$ do
    $K_j \leftarrow \{c_j\}$
    for $i = 2 \ldots T$ do
        Compute $x_j$ based on $K_j$ (Eq. 2)
        $w = \arg \max_{w \notin K_j} \text{sim}(s_w, x_j)$
        Compute $x'_j$ based on $K_j \oplus \{w\}$
        // consistency check
        if $x'_j$ changes the words in $K_j$ then
            break
        else
            $K_j \leftarrow K_j \oplus \{w\}$

// document rep. estimation
for $i = 1 \ldots n$ do
    for attention mechanism $m \in \mathcal{M}$ do
        Rank $D_{i,j}$ according to $m$
        $r_{m,j} \leftarrow$ the rank of $D_{i,j}$
        Rank $D_{i,j}$ according to $\prod_m r_{m,j}$
        $r_j \leftarrow$ the final rank
        $a_j \leftarrow 1/r_j$
        $E_i \leftarrow \sum_j a_j x_{i,j} / \sum_j a_j$
```
M2: Document-Class Alignment & M3: Text Classifier Train

- **M2: Document-Class Alignment**
  - each document is assigned to its nearest class
  - Gaussian Mixture Model (GMM) clustering
  
  \[ L_i = \arg \max_c \cos(E_i, x_c) \]

- **M3: Text Classifier Training**
  - select most confident samples to train a text classifier (BERT) using the pseudo labels
Experiments

Table 2: Evaluations of Compared Methods and X-Class. Both micro-/macro-F\textsubscript{1} scores are reported. WeSTClass and ConWea consume at least 3 seed words per class. Supervised provides a kind of upper bound. We are not able to re-run WeSTClass and ConWea on DBpedia due to the large size.

<table>
<thead>
<tr>
<th>Model</th>
<th>AGNews</th>
<th>20News</th>
<th>NYT-Small</th>
<th>NYT-Topic</th>
<th>NYT-Location</th>
<th>Yelp</th>
<th>DBpedia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>93.99/93.99</td>
<td>96.45/96.42</td>
<td>97.95/95.46</td>
<td>94.29/89.90</td>
<td>95.99/94.99</td>
<td>95.7/95.7</td>
<td>98.96/98.96</td>
</tr>
<tr>
<td>WeSTClass</td>
<td>82.3/82.1</td>
<td>71.28/69.90</td>
<td>91.2/83.7</td>
<td>68.26/57.02</td>
<td>63.15/53.22</td>
<td>81.6/81.6</td>
<td>81.1/ N/A</td>
</tr>
<tr>
<td>ConWea</td>
<td>74.6/74.2</td>
<td>75.73/73.26</td>
<td>95.23/90.79</td>
<td><strong>81.67/71.54</strong></td>
<td>85.31/83.81</td>
<td>71.4/71.2</td>
<td>N/A</td>
</tr>
<tr>
<td>LOTClass</td>
<td><strong>86.89/86.82</strong></td>
<td>73.78/72.53</td>
<td>78.12/56.05</td>
<td>67.11/43.58</td>
<td>58.49/58.96</td>
<td>87.75/87.68</td>
<td>86.66/85.98</td>
</tr>
<tr>
<td>X-Class</td>
<td>84.8/84.65</td>
<td><strong>81.36/80.6</strong></td>
<td><strong>96.67/92.98</strong></td>
<td>80.6/69.92</td>
<td><strong>90.5/89.81</strong></td>
<td><strong>88.36/88.32</strong></td>
<td><strong>91.33/91.14</strong></td>
</tr>
<tr>
<td>X-Class-Rep</td>
<td>77.92/77.03</td>
<td>75.14/73.24</td>
<td>92.13/83.94</td>
<td>77.85/65.38</td>
<td>86.7/87.36</td>
<td>77.87/77.05</td>
<td>74.06/71.75</td>
</tr>
<tr>
<td>X-Class-Align</td>
<td>83.1/83.05</td>
<td>79.28/78.62</td>
<td>96.34/92.08</td>
<td>79.64/67.85</td>
<td>88.58/88.02</td>
<td>87.16/87.1</td>
<td>87.37/87.28</td>
</tr>
</tbody>
</table>
Experiments

(a) Our Class-Oriented Document Representations  (b) Simple Average of BERT Representations

Figure 5: T-SNE Visualizations of Representations. From left to right: NYT-Topics, NYT-Locations, Yelp.