Multi-Domain Text Classification

Domains are not born equal

MAN for MDTC

\[ J_p \quad \lambda J_p \quad J_u \]

\[ J_p(d_i \in \Delta_k) \]

\[ \text{Domain Discriminator} \]

\[ \text{Test Classifier} \]

\[ \text{Shared Feature Extractor} \]

\[ \text{Domain-Specific Feature Extractor} \]

The non-cooperative coupling of \( F_d \) and \( D \) form a Multinomial Adversarial Network. \( F_d \) attempts to identify the domain of a sample using the shared features. \( F_d \) learns domain-invariant features by learning to stop leaking domain information to \( D \).

MAN Theory

**Summary:** We show that MAN, as a versatile machine learning framework, directly minimizes the \( f \)-divergence among multiple distributions.

\( f \)-divergence is a family of metrics measuring the difference between probability distributions. Many common divergences, such as KL-divergence, total variation divergence, are special cases of \( f \)-divergence.

Nowozin et al. (2016) proved that standard adversarial nets are minimizers of various \( f \)-divergence metrics between two distributions, depending on the choice of loss function.

MAN is hence a generalization of the impactful (bimodal) adversarial networks to multiple distributions.

Let the distribution of the shared features \( f \) for instances in each domain \( d_i \in \Delta \) be:

\[ P_d(d_i) = \prod_{j=1}^{N} P_{d_j}(d_i) \]

We consider two MAN variants with the NLL and L2 loss, respectively:

\[ J_{\text{NLL}} = \sum_{i=1}^{N} \sum_{j=1}^{D} P_{d_j}(d_i) \log(P_{d_j}(d_i)) \]

\[ J_{\text{L2}} = \sum_{i=1}^{N} \sum_{j=1}^{D} (P_{d_j}(d_i))^2 - 1 \]

**Theorem 1.** If \( D \) is trained to its optimality, if \( D \) adopts the NLL loss:

\[ J_{\text{NLL}} = -\log(\pi) = -\log N + \sum_{i=1}^{N} \log(\pi_d(d_i)) \]

where \( \pi_d(d_i) \) is the generalized Jensen-Shannon Divergence (Lim, 1991) defined as the average Kullback-Leibler divergence of each \( P_d(d_i) \) to the centroid \( \pi \) (Aslam and Paul, 2017).

**Theorem 2.** If \( D \) uses the L2 loss:

\[ J_{\text{L2}} = \sum_{i=1}^{N} \sum_{j=1}^{D} (P_{d_j}(d_i))^2 \]

**Corollary 1.** The optimum of \( J_{\text{NLL}} \) is \(-\log N\) when using NLL loss, and 0 for the L2 loss. The optimum value above is achieved if and only if \( P_0 = P_1 = \cdots = P_N = \pi \) for either loss.

MAN Glossary

**D** The set of all labeled domains which have some annotated data

**A** The set of all domains, \( A = \{ \Delta_d \} \)

**X** The labeled data for domain \( d \in A \)

**U** The unlabeled data for domain \( d \in A \)

**F_d** The shared feature extractor that extracts domain-invariant features

**C_d** The domain-specific feature extractor that extracts domain-specific features

**F** The adversarial domain-discriminator

**J_p** The cost minimization

**J_u** The cost maximization

**J_d** The domain cost of \( F_d \) that is uncorrelated to \( J_p \)

**\lambda** A hyperparameter synchronizing the training of \( D \) and the rest of MAN

**NLL** The NLL criterion with the Negative Log Likelihood loss

**L2** The NLL criterion with the Least Square loss

**Experiments**

**MDTC Experiments**

**Amazon Dataset (Table 1)**

- Widely adopted
- 2000 samples/domain
- 5-fold cross-validation
- 4 domains
- Preprocessed to bag-of-words features (no raw text, no word order information)

**FDU-MTL Dataset (Table 3)**

- Less reported results
- ~2000 samples/domain
- Pre-split train/dev/test sets
- 16 domains
- Original texts available

**Evaluation Metrics**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN</td>
<td>0.95</td>
<td>0.89</td>
<td>0.91</td>
</tr>
<tr>
<td>MAN</td>
<td>0.97</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>ASP-MAN</td>
<td>0.98</td>
<td>0.94</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 1: MDTC results on the Amazon dataset. Models in bold are ours while the performance of the rest is taken from Wu and Huang (2015). Numbers in parentheses indicate standard errors, calculated based on 5 runs. Bold numbers indicate the highest performance in each domain, and \( x \) shows statistical significance (\( p < 0.05 \)) over CMSC under a one-sample \( t \)-Test.

**Experiments on Unlabeled Domains**

**Baselines:** Multi-Source Domain Adaptation Methods

**Dataset:** Amazon (4 domains)

**3 labeled (source) domains**

**1 unlabeled (target) domain**

**MAN achieves SToA performance**

**MAN also has the potential to handle >1 unlabeled domains**

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
<th>F1-Score</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSLDA</td>
<td>0.85</td>
<td>0.90</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>DAN</td>
<td>0.88</td>
<td>0.92</td>
<td>0.90</td>
<td>0.92</td>
</tr>
<tr>
<td>MAN</td>
<td>0.92</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>ASP-MAN</td>
<td>0.96</td>
<td>0.97</td>
<td>0.98</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table 2: Results on unlabeled domains. Models in bold are ours while the rest is taken from Zhao et al. (2017). Highest domain performance is shown in bold.