Background

- While the Question Answering (QA) task is a promising application of NLP, its ability to generalize to new datasets remains a challenge.
- Models tend to overfit to specific datasets, or domains, they are trained on, decreasing their utility in real-world applications.
- In the past, adversarial training has been applied to produce domain-agnostic question-answering. See figure 1 below.

Task: Partial Domain Independence

![Image: Figure 1: Training procedure for learning domain-agnostic feature representations. The discriminator is trained to predict the domain of the dataset based on the output (CLS) token. The model classifier predicts the appropriate answer while fooling the discriminator. Taken from Lee et al. 2019.]

Adding an Adversarial Component to the Baseline

- We explore creating partially domain-invariant models that improve performance of the model while remaining generalizable.
- Our final loss function for the QA model can be written as: \( L = L_{\text{basel}} + \lambda L_{\text{adv}} \)
- The influence of the discriminator loss is set by the hyperparameter \( \lambda \).
- \( L_{\text{basel}} \) represents classification loss while \( L_{\text{adv}} \) is discriminator loss.

Approach

- The adversarial network was implemented using different values of lambda in the loss function: \( L_{\text{adv}} = \lambda L_{\text{adv}} \).
- A lambda value of 0.01 led to the best performance metrics, with an EM value of 31.94 and an F1 score of 49.321.
- This value was used in subsequent model trainings.

Partially Domain-Invariant Models

- Model performance was assessed when features were trained to be partially independent of the domain.
- In each case, a component of the feature vector (\( \beta \)) was trained on the discriminator while the remaining component was directly passed to the classifier.

Model Refinement with Wasserstein Distance

- We explored model improvements by replacing the Kullback-Leiber (KL) divergence with a Wasserstein distance measure to adversarially train the discriminator function.
- At a high level, the Wasserstein distance is a distance metric between two probability distributions, defined as: \( W(P_\text{source}, P_\text{target}) = \inf_{f \in \Phi} \mathbb{E}_{x \sim P_\text{source}} \| x - f(x) \|_1 \)
- \( f(P, P_\text{target}) \) is the set of all joint distributions over \( x \) and \( y \) such that the marginal distributions are equal to \( P \) and \( P_\text{target} \).
- In this case, the predicted domain from the discriminator is representative of the source domain and a uniform distribution is the target domain.

Handling Class Imbalance with Focal Loss

- Focal loss was implemented to handle imbalance in predictions caused by class imbalance in the training set.
- It adds a factor \((1-p)^\gamma\) to the standard cross entropy term, allowing the loss function to apply more focus to misclassified examples.

Final Results

<table>
<thead>
<tr>
<th>Gamma</th>
<th>Alpha</th>
<th>Exact Match</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>0</td>
<td>30.63</td>
<td>46.69</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>30.63</td>
<td>47.41</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>31.68</td>
<td>47.35</td>
</tr>
<tr>
<td>2.05</td>
<td>0.25</td>
<td>33.77</td>
<td>48.92</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>31.94</td>
<td>47.01</td>
</tr>
</tbody>
</table>

In concordance with the results from developers of focal loss (Lin et al. 2018), a gamma value of 2.0 and alpha value of 0.25 provided the best performance.

Wasserstein Distance

<table>
<thead>
<tr>
<th>Lambda</th>
<th>Adversarial Loss Training</th>
<th>Sample Type</th>
<th>Exact Match</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>Wasserstein</td>
<td>Weighted</td>
<td>31.94</td>
<td>48.49</td>
</tr>
<tr>
<td>0.05</td>
<td>Wasserstein</td>
<td>Random</td>
<td>32.72</td>
<td>49.24</td>
</tr>
<tr>
<td>0.01</td>
<td>KL-Oswieger</td>
<td>Weighted</td>
<td>29.32</td>
<td>43.78</td>
</tr>
<tr>
<td>0.01</td>
<td>KL-Oswieger</td>
<td>Random</td>
<td>35.079</td>
<td>49.321</td>
</tr>
</tbody>
</table>

- While KL-Oswieger demonstrated optimal performance on the models tested, a lambda value of 0.05 provided best performance on models implemented with Wasserstein distance.

Combining Focal Loss and Wasserstein Distance

- When both techniques are combined (with hyperparameters \( \lambda = 0.01 \), \( \alpha = 0.25 \), \( \beta = 0.95 \), \( \gamma = 2.0 \)), we achieve our best performance, with F1=51.16 and EM=35.08 on the dev set and an F1=60.069 and EM=41.789 on the test set.

Summary

- Compared to our baseline model trained without an adversarial component, adding the discriminator improved performance in terms of F1-Score and Exact Match (EM). Developing features with partial domain independence also improved the model’s performance on unseen data.
- While our dataset was heavily imbalanced, it remains unclear whether focal loss improved overall performance.
- While several combinations of hyperparameters were tested, a more extensive and organized hyperparameter search needs to be conducted to make conclusions on the utility of Wasserstein distance and focal loss.

References