Out-of-domain Detection for Sequence Data

Patrick Hayes
Mahdi Behroozikhah
Samuel Sunarjo
What is Out-of-domain Detection?

- Detect data that differs substantially from the dataset which the model is trained on.
- Examples
  - MNIST vs. CIFAR
  - Wall Street Journal vs. Twitter
  - Different newsgroup subjects
Why Out-of-domain Detection?

- ML models tend to fail when the training and test distributions are different.
- Moreover, the models are unable to indicate when they are likely mistaken.
  - Hinders the adoption of ML in technologies due to AI safety concern.
- Anomaly detection.
- Network security vulnerabilities
  - One Pixel Attack for Fooling Deep Neural Networks.
  - If the training data is supposed to be private.
- An important task in computer vision is feature extraction which is highly susceptible to adversarial attacks and domain drift. In natural language processing there is much less feature extraction because the words or n-grams are used as features.
Adversarial Examples

“panda” 57.7% confidence

+ $\epsilon$

= 99.3% confidence

“gibbon”
Domain Shift

Training data and test set are of summer driving conditions -> Get **95% accuracy** on pedestrian detection

Show example of pedestrian in the winter -> **Fails miserably**
Prior Work

A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks

Dan Hendrycks*
University of California, Berkeley
hendrycks@berkeley.edu

Kevin Gimpel
Toyota Technological Institute at Chicago
kgimpel@ttic.edu

- Proposed a simple baseline that utilizes probabilities from softmax distributions for detection.
- Addressed tasks in a few domains, Computer Vision, Natural Language Processing, Audio Speech Recognition
Problem Formulation & Evaluation

- **Success and Error prediction**
  - Can we predict whether a trained classifier will predict **correctly or incorrectly** on a test example.

- **In- and out-of-distribution detection**
  - Can we predict whether a test example is from a **different or same** distribution from the training data.

- **AUROC**
  - The ROC curve plots the true positive rate vs. false positive rate.
  - Probability that a positive example has a greater detector score/value than a negative example.
  - A random detector corresponds to a 50% AUROC, and a “perfect” classifier corresponds to 100%.

- **AUPR**
  - The PR curve plots the precision and recall.
  - Baseline detector has an AUPR approximately equal to the precision and a “perfect” classifier has an AUPR of 100%.
Baseline - Method

- Retrieve the maximum/predicted class probability from a **softmax distribution** and detect whether an example is erroneously classified or out-of-distribution.

- Success and Error prediction
  - Separate correctly and incorrectly classified test set examples and compute the softmax of the predicted/max class.
  - In order to obtain ROC and PR for erroneous test examples, label incorrectly classified examples as positive and take the negatives of the softmax of the predicted classes as the scores.

- In- and out-of-distribution detection
  - For “In”: treat the in-distribution, correctly classified test set examples as positive and use the softmax of the predicted class as the score
  - For “Out”: treat the out-of-distribution examples as positive and similarly use the negative of the softmax.
Our Goals

- Reproduce baseline results.
- Propose new approaches for detection in sequence and NLP tasks.
- NLP Tasks
  - Sentiment Analysis
  - Categorization
  - Part-of-speech (POS) tagging
- Use the same metrics to evaluate our performance.
NLP Task - Sentiment Analysis

- Binary Sentiment Classification
- Dataset
  - In-Domain:
    - IMDB Dataset - Maas et al(2011)
      - “… I haven't seen the first Nemesis film, but I did check the info out of it and I here by say…”
  - Out-of-Domain:
      - “a movie with a real anarchic flair”
      - “it has a 3x optical zoom , which is average for these cameras”
      - “don't buy this player”
NLP Task - Categorization

- Predict the subject of the text.
- Dataset
  - In-Domain:
      - Political: livesey writes now along comes keith schneider and says here objective moral system...
      - Copper: mining corp said cutting its copper cathode price cent cents effective immediately.
  - Out-of-Domain:
NLP Task - Part-of-speech (POS) tagging

- Identify the part of speech for each word in a text, based on both its definition and context.
- Some words can represent more than one part of speech at different times.
- Dataset
  - In-Domain:
    - **Twitter Dataset** - Gimpel et al. (2011)
      - 25 tags.
      - I predict I won’t win a single game I bet on.
      - O V O V V D A N O V P
      - O: Pronoun, D: Determiner, P: Pre/Postposition
  - Out-of-Domain:
    - **Wall Street Journal** - Marcus et al. (1993)
Our Method and Implementation

For an out of domain example the distance is random.

Confidence is the distance from the decision boundary.

A random out of domain example is more likely to have a low confidence.

Confidence is represented by a mixture of gaussians.
Our Method and Implementation

- Train a distribution approximator on each output layer of the models for each task.
  - Gaussian Mixture Model (GMM)
- Train GMM on output of FastText; on logits on the final FC layer for the task.
- Train GMM on output of each layer (before non-linearity) in the MLP model for twitter (PoS).
- Train GMM on output of each conv layer (before non-linearity) of the CNN for texts.
### Result - Success and Error Prediction

<table>
<thead>
<tr>
<th>Dataset</th>
<th>AUROC / Base</th>
<th>AUPR Succ / Base</th>
<th>GMM AUPR Succ / Base</th>
<th>AUPR Err / Base</th>
<th>GMM AUPR Err / Base</th>
<th>Pred. Prob Wrong</th>
<th>Pred. Prob Right</th>
<th>Test Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB</td>
<td>82/50</td>
<td>97/88</td>
<td>97/88</td>
<td>37/12</td>
<td>78/12</td>
<td>71</td>
<td>89</td>
<td>11.6</td>
</tr>
<tr>
<td>15 Newsgroups</td>
<td>89/50</td>
<td>99/93</td>
<td>87/93</td>
<td>46/7.3</td>
<td>95/7.3</td>
<td>44</td>
<td>82</td>
<td>8</td>
</tr>
<tr>
<td>Reuters 52</td>
<td>95/50</td>
<td>99/89</td>
<td>96/89</td>
<td>74/11</td>
<td>81/11</td>
<td>37</td>
<td>93</td>
<td>11</td>
</tr>
<tr>
<td>Twitter</td>
<td>89/50</td>
<td>98/87</td>
<td>90/87</td>
<td>55/13</td>
<td>84/13</td>
<td>71</td>
<td>96</td>
<td>13</td>
</tr>
</tbody>
</table>

*“Base”: values that a random detector would obtain.

*“Success” or “Succ”: correctly classified.  
*“Error” or “Err”: incorrectly classified.

“Pred. Prob Wrong”: the mean predicted class probability of incorrectly classified examples.
Result - In/Out-of-domain Detection

<table>
<thead>
<tr>
<th>In-Domain / Out-Domain</th>
<th>AUROC / Base</th>
<th>AUPR In/Base</th>
<th>GMM AUPR In/Base</th>
<th>AUPR Out/Base</th>
<th>GMM AUPR Out/Base</th>
<th>Pred. Prob Wrong</th>
</tr>
</thead>
<tbody>
<tr>
<td>IMDB/Customer</td>
<td>95/50</td>
<td>99/89</td>
<td>99/89</td>
<td>60/11</td>
<td>48/11</td>
<td>60</td>
</tr>
<tr>
<td>IMDB/Movie</td>
<td>94/50</td>
<td>98/72</td>
<td>97/72</td>
<td>80/28</td>
<td>72/28</td>
<td>61</td>
</tr>
<tr>
<td>15/5 Newsgroups</td>
<td>75/50</td>
<td>52/33</td>
<td>32/33</td>
<td>86/66</td>
<td>63/66</td>
<td>58</td>
</tr>
<tr>
<td>Reuters40/Reuters12</td>
<td>80/50</td>
<td>99/86</td>
<td>96/86</td>
<td>80/13</td>
<td>30/13</td>
<td>33</td>
</tr>
</tbody>
</table>

*“Base”: values that a random detector would obtain.

“Pred. Prob Wrong”: the mean predicted class probability of wrongly classified examples.
Conclusion so far

- Gaussian Mixture Models can be difficult to train. Future work is needed to explore the best way to set hyper parameters and test whether modifications to the traditional GMM could be made to improve learning.
- In FastText the feature extraction is done using word embeddings. If a word has not been seen before then it is out of domain. Our method works better for images where the feature extraction is done using many neural network layers.
Potential Next Steps

- Apply an out-of-domain mechanism to the BERT language model and other architectures.
- Use other distribution approximator for OOD (instead of just GMM) and compare them.
  - BayesianGaussianMixture
- More experiments on PoS task with additional dataset. One of the datasets in the reference paper has a license fee of $3k, which we could not use to conduct OOD experiments with the twitter dataset.
Thank you!