Overview / Motivation
- Robustness is among the most important attributes in machine learning models: many, perhaps even most, tasks have very little training data
- Question-Answering (QA) systems are often not reliable on out-of-domain (OOD) inputs, even while surpassing human-level performance on in-domain (ID) inputs (i.e., SQuAD dataset)
- Hinders their deployment in real-world settings

Data Augmentations
- **Data Mixing**: Splice ID and OOD context paragraphs together.
- **Selective Masking**[1]: Randomly mask part of training input and refill using a BERT LM finetuned on OOD data.
- **Easy Data Augmentation (EDA)**[2]: A set of 4 data augmentation techniques for text: random synonym replacement, random insertion, random swap, and random deletion
- **Back-Translation**[3]: Back-translate the question and non-answer sentences in the context using a pivot language (French) to generate augmented examples with similar phrasing
- **Self-Supervised**[4]: Generate pseudo-tasks using the OOD test dataset such that the model can learn the distribution of this domain, similar to Test-Time Training approach in vision domain

Data
- Datapoint is question and context pair. Label is a span of text from context (start/end indices)
- In domain (ID) and out of domain (OOD)
  - ID: SQuAD, NewsQA, Natural Questions
  - OOD: DuoRC, RelationExtraction, RACE

Setup
- All experiments fine-tune the Hugging Face DistilBERT for 3 epochs, using AdamW with learning rate 3E-5.
- Back-Translation - used Helsinki-NLP pre-trained MT model, trained on MarianNMT and OPUS.
- Selective Masking - non-answer tokens masked with prob=0.15. DistilBERT LM finetuned on OOD data.

Results
- While baseline has strong performance on the ID data (70.71 F1, 54.69 EM), it struggles on OOD.
- Original goal was to explore the data mixing approach deeply, to expose the model with both ID and OOD domain data during training time. Exploring other approaches has currently out performed data mixing.
- EDA, while simple, has surprisingly performed well in improving OOD robustness.
- Combination of self-supervised and back-translation training has worked best so far

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>EM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>47.78</td>
<td>31.68</td>
</tr>
<tr>
<td>Data Mixing</td>
<td>41.72</td>
<td>25.13</td>
</tr>
<tr>
<td>Selective Masking</td>
<td>47.73</td>
<td>32.98</td>
</tr>
<tr>
<td>Back-Translation</td>
<td>48.72</td>
<td>34.82</td>
</tr>
<tr>
<td>EDA</td>
<td>49.75</td>
<td>36.65</td>
</tr>
<tr>
<td>Self-Supervised</td>
<td>51.87</td>
<td>37.72</td>
</tr>
<tr>
<td><strong>Self-Supervised + Back Translation</strong></td>
<td><strong>52.91</strong></td>
<td><strong>38.75</strong></td>
</tr>
</tbody>
</table>

Figure 1. OOD validation set performance (F1, EM)

Conclusion/Discussion
- Data augmentation helps improve the OOD robustness of the model via improving the data distribution support of the training dataset
- Selective masking did not work as well as we expected, and we wish to further explore this.
- Data augmentation techniques improve OOD performance across the board, although our original idea of mixup has not
- Self-supervised learning is the most performant, as it allows us to capture the features unique to OOD data distribution and thus improve OOD robustness

Future Work
- Currently, EDA performs better than back-translation
- Thus, we plan to combine EDA with self-supervised learning to see if that further improves the OOD performance
- Re-explore mixup[5] using a technique called LINDA[6], to interpolate between ID and OOD as data augmentation
- We currently only use selective masking to generate on ID data, and would be interested to explore also using OOD to generate new samples

References/Prior Work