INTRODUCTION
Question Answering (QA) is a critical task for NLP applications such as conversational agents and search engines in which generalization to new domains is highly desirable.

BACKGROUND
SOTA QA models often fail to generalize to new domains without significant fine-tuning. We aim to build a robust QA model using adversarial learning approach. Lee et.al. achieved improved performance in terms of EM and F1 using Adversarial approach on MRQA Shared Task 2019.

DATASET
3 In-Domain Datasets : SQuAD, NewsQA, Natural Questions
3 Out-of-Domain datasets : DuoRC, RACE, RelationExtraction

METRICS
F1 score : the harmonic mean of precision and recall
Exact Match : a binary measure (i.e. true/false) of whether the system output matches the ground truth answer exactly

METHODS
Our Adversarial Training approach consists of:
- **Generator Model**: pre-trained DistilBERT
- **Discriminator Model**: 3-layer MLP

$$\mathcal{L}_G = \mathcal{L}_{QA} + \lambda \mathcal{L}_{adv}$$

EXPERIMENTS
For our adversarial experiments, we tuned:
- Lambda (i.e. weight of adversarial loss)
- Dropout
- Hidden size of Discriminator

BEST MODEL
Best performance on Out-of-Domain Validation Set for:
- Lambda = 0.01
- Dropout = 0.2
- Hidden Size of Discriminator = 768

RESULTS

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Model</th>
<th>Our best model</th>
<th>baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>indomain-val</td>
<td>F1</td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>70.68</td>
<td>54.25</td>
<td>70.43</td>
</tr>
<tr>
<td>oodomain-val-all</td>
<td>49.75</td>
<td>32.98</td>
<td>49.0</td>
</tr>
</tbody>
</table>

ANALYSIS
- Large Dropout and small Lambda boosts discriminator and forces generator to learn domain in-variant features.
- Score improvement on in-domain dataset doesn’t improve score for all oo-domain datasets in general.

CONCLUSIONS
Adversarial Training helps the QA model generalise to out-of-domain datasets, and shows improved performance over the baseline on oo-domain dataset for F1 score by 0.75.

References: