In this project, we explore building and improving a neural model for solving the question-answering tasks defined by SQuAD 2.0[1], in order to contribute to this popular NLP research topic, and to better understand techniques of applying neural models to NLP tasks in general.

**Data/Task**

**Task & Dataset:** The Stanford Question Answering Dataset (SQuAD)[1]
- **Input:**
  - Context (a text passage)
  - Question (reading comprehension)
- **Output:**
  - Answer to the question (plain text)
  - AvNA (Answer vs. No-Answer)

**Evaluation Metrics:**
- Exact-Match (EM) Score
  \[ \text{EM} = \frac{\text{exact matching answers}}{\text{total evaluated questions}} \times 100 \]
- F1 Score
  \[ P = \frac{\text{true positives}}{\text{true positives + false positives}} \]
  \[ R = \frac{\text{true positives}}{\text{true positives + false negatives}} \]
  \[ F1 = 2 \cdot \frac{PR}{P + R} \]
- AvNA (Answer vs. No-Answer)

**Q&A Example:**

**Question:** Economy, Energy and Tourism is one of the what?

**Context:** Subject Committees are established at the beginning of each parliamentary session, and again the members on each committee reflect the balance of parties across Parliament. Typically each committee corresponds with one (or more) of the departments (or ministries) of the Scottish Government. The current Subject Committees in the fourth Session are: Economy, Energy and Tourism; Education and Culture; Health and Sport; Justice; Local Government and Regeneration; Rural Affairs, Climate Change and Environment; Welfare Reform; and Infrastructure and Capital Investment.

**Correct Answer:** current Subject Committees

**Prediction (Baseline):** N/A

**Prediction (Single Transformer):** The current Subject Committees

**Prediction (Char-Embedding):** current Subject Committees

**Prediction (Res-BiDAF, single block):** current Subject Committees

**Prediction (Res-BiDAF, 2 blocks):** current Subject Committees in the fourth Session

**Models with major performance improvements:**
- Transformer[2] (Single / Double)
- Char-level Embedding
- Res-BiDAF (Single Block)
- Res-BiDAF (Double Block)

**Self-Attended Residual Bi-Directional Attention Flow (Res-BiDAF):**
- Char-level embedding concat with word embeddings
- Gated Residual Blocks
  \[ \text{LSTM + BiDAF + ReLU Gate, summed with pass-through} \]

**Post-prediction TF-IDF filtering**
- TF-IDF: Relevance metric, higher score for rare word appearances
- We length-normalized TF-IDF to measure prediction relevance
- Manually mark “no answer” if a prediction has low TF-IDF score (i.e. irrelevant / not interesting)

**Analysis**

**Alternative models brought significant performance improvements**

<table>
<thead>
<tr>
<th>Model</th>
<th>F1 Score</th>
<th>EM Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>58.98</td>
<td>55.81</td>
</tr>
<tr>
<td>Single Transformer</td>
<td>59.70</td>
<td>57.18</td>
</tr>
<tr>
<td>Char-Embedding</td>
<td>64.36</td>
<td>60.98</td>
</tr>
<tr>
<td>Res-BiDAF (single block)</td>
<td>65.40</td>
<td>62.06</td>
</tr>
<tr>
<td>Res-BiDAF (2 blocks)</td>
<td>67.69</td>
<td>64.48</td>
</tr>
</tbody>
</table>

**Dev set AvNA / EM / F1 score curves**

**Non-PCE Test Leaderboard submissions**
- Res-BiDAF (single block) EM: 58.698 F1: 62.333
- Res-BiDAF (2 blocks) EM: 62.992 F1: 66.370

**TF-IDF post-filtering failed to produce satisfactory results**

**Conclusion / Future Work**
- Our proposed Res-BiDAF reached significant performance improvements over baseline non-PCE model
- Still far away from catching up with PCE models (ELMo & BERT)
- Potential future work:
  - Train Res-BiDAF with higher # of Gated Residual Blocks
  - Combine Res-BiDAF with other non-PCE techniques
  - Experiment applying Res-BiDAF to PCE models

---