**Problem**

**Objective:** Given a query and context, QANet aims to answer the query with a selection from its corresponding context or "NA" if no answer exists.

**Motivation:** QANet achieved state-of-the-art question-answer performance on SQuAD 1.1 (2018), where all queries had answers.

- We implemented QANet, testing its performance on question answering for SQuAD 2.0, which includes "no answers" to some questions.

**Evaluation Metrics:**
- **F1 Score:** Harmonic mean of precision & recall. 
  \((2 \times \text{precision} \times \text{recall}) / (\text{precision} + \text{recall})\)
- **EM Score:** Binary exact match for ground truth answers (true/false)

**Top Model:** Ensemble **BiDAF**(200-dim)+*(64-dim)+QANet(w/ 4-heads) **EM/F1 = 63.02/66.4**

---

**Experiments**

**Model** | **Dev EM/F1** | **Test EM/F1**
--- | --- | ---
BiDAF: Baseline | 57.87/60 | 58.11/60
BiDAF: 64-dim char embeds | 60.66/64 | 61.93/65
BiDAF: 200-dim char embeds | 62.46/65 | 62.19/65
QANet: 2 heads, 7 blocks | 56.26/60 | 56.54/60
QANet: 4 heads, 5 blocks | 57.18/61 | 57.46/61
QANet: 8 heads, 7 blocks | 61.2/65 | 61.49/65
Ensemble (200-dim, 48.2 head) | 63.02/66 | 63.31/66

Short orange=BiDAF (baseline), blue=BiDAF(64-dim char embeds), red=BiDAF(200-dim char embeds), Long orange=QANet(4-heads), cyan=QANet(2-heads)

---

**Analysis**

- **Depth:** More encoding blocks did not significantly increase F1 scores but did significantly increase training time. If discovered earlier, we would have decrease depth and increased our batch size, allowing for more experimentation.

- **Inconsistent Hyperparameters:** Due to memory limitations, we used different hyperparameters (e.g., quantities of heads & encoder blocks affected permissible batch sizes). Since we could not keep hyperparameters consistent, we cannot draw strong conclusions about their influence.

- **Common Errors:** Via manual inspection, we found that our model struggles with answers containing numbers.

---

**Conclusions**

- **Results non-transferable to different dataset:** Original QANet needs significant modifications to perform as well on SQuAD 2.0 as it did on version 1.1 (in speed & accuracy).

- **More layers/heads are not always better:** Diminishing returns exist for layer depth/heads number to performance payoff (less encoder blocks achieved nearly the same F1 scores in faster training & 8 heads was worse than 4 heads).

- **Faster?** Yu et al. found QANet faster at training and inference than RNN models. We did not find QANet faster nor more accurate than BiDAF, but results likely differ with better hardware.

- **Future Work:**
  - Given we used a subset of SQuAD 2.0, training on a larger subset with 8-headed attention and a larger batch size of 32 would give a better comparison with Yu et al.’s results.
  - Several trainings with different seeds gave us significantly different results. Thus, training at different seeds and ensemble models together might be useful.
  - Integrate Transformer-XL to get longer range dependencies.

---

**References**