Problem
- The goal is to develop a Question-Answering Model, which takes a Question and Paragraph as inputs, and attempts to answer the question as correctly as possible - providing a measure for how well the model can understand "text".
- The baseline Model is based on BiDirectional Attention Flow (BIDAF) Architecture.
- We implemented QANet Architecture, which uses Convolution and Self-Attention to replace the Sequential Recurrent Networks from the baseline Model.

Data
- Stanford Question Answering Dataset (SQuAD) v2.0
- Around 150K Questions.
- More than half the questions can’t be answered using the paragraph.
- Data Split into: ~ 90.6% Train, 4.3% Dev, 4.2% Test.

Methods
Character Embedding: Word-level embeddings do not address morphemes, misspelled or out-of-vocabulary words. We add character-level embedding to enhance input representation.

Token Features: Factual Q&As benefit from input features such as Part-of-Speech, Named-Entity Recognition, and Frequency.

Data Augmentation: Techniques used were:
- Back-Translation using different Languages (French, Chinese, Spanish, Hindi) - to rephrase the text and introduce diversity.
- Synonym Replacement - to introduce new vocabulary into the text.
- Basic sanity checks were added to validate the augmented text: such as detecting changes in the Named Entities.

Approach
Implemented QANet, a transformer-like model which has higher speed and accuracy over BIDAF.

Results/Analysis
- Single models
  - BIDAF: baseline
    - F1: 62.70, EM 57.99, AR:NA
  - BIDAF: char_emb
    - F1: 63.54, EM 60.07, AR:NA
  - BIDAF: char_emb, token_free
    - F1: 64.69, EM 62.70, AR:NA
- QANet
  - QConv. Head, Rescale, 2kD, emb: 66.02, EM 62.71, AR:NA
  - QConv. Head, 2kD, 30kD, emb: 68.51, EM 64.93, AR:NA
- QANet**: QAConv. Head, 3kD, emb: 67.98, EM 64.21, AR:NA
- QANet**: QAConv. Head, 3kD, emb: 69.44, EM 65.89, AR:NA
- QANet**: QAConv. Head, 3kD, emb: 69.87, EM 66.12, AR:NA
- QANet**: QAConv. Head, 3kD, emb: 70.77, EM 67.41, AR:NA
- QANet**: QAConv. Head, 3kD, emb: 71.54, EM 68.48, AR:NA
- QANet**: QAConv. Head, 3kD, emb: 72.55, EM 69.65, AR:NA
- QANet**: QAConv. Head, 3kD, emb: 73.49, EM 70.89, AR:NA

- Ensemble models
  - QANet ensemble: average prediction
    - F1: 71.73, EM 68.73, AR:NA
  - QANet ensemble: majority voting
    - F1: 71.4, EM 68.55, AR:NA
  - QANet-BIDAF ensemble: majority voting
    - F1: 73.3, EM 69.7, AR:NA

- Basic QANet model outperformed BIDAF achieving 68.51/64.93 F1/EM score. Complexity was gradually added, in order to evaluate the importance of each element on the performance.
  - No benefit was seen in increasing the Model Encoder Stack from 5 to 7.
  - A big improvement (+2.5 F1 Score) was seen by increasing Attention heads from 1 to 8, Hidden size from 96 to 128.
- Character Embedding and Token features were the most important enhancements on the architecture giving +1.1 F1 score gain each.
- Data Augmentation was effective in diversifying the input data-set for both Questions and Paragraphs.
- Ensemble gave a better prediction than any stand-alone model. Average probability performed better than majority voting.

Conclusion
- QANet out-performed BIDAF.
  - All the architectural changes and fine-tuning of parameters ended up with the highest scores of:
    - 69.44/65.89 F1/EM score on the dev set with single-model
    - 72.55/69.65 F1/EM on the dev set with ensemble
    - 69.73/67.22 F1/EM score on the hidden test set
- The most-common mistake is answering un-answerable question. Adding a separate head for no-answer may help.

References

Output Layer was further enhanced with conditioning the pwx token on pwx. Two Methods were tested:

- **Method A**: $X_{pwx} = \max_{i,j} \{M_{i,j}\} 
  $X_{pwx} = \{W_{i} \times [M_{i}]\} $\max_{i,j} = \max_{i,j}(W_{i} \times [M_{i}]\}$

- **Method B**: $X_{pwx} = F_{bilinear}(W_{i}, W_{x}) 
  $\max_{i,j} = \max_{i,j}(W_{i} \times [X_{i}]\)$

Conclusion
- QANet out-performed BIDAF.
  - All the architectural changes and fine-tuning of parameters ended up with the highest scores of:
    - 69.44/65.89 F1/EM score on the dev set with single-model
    - 72.55/69.65 F1/EM on the dev set with ensemble
    - 69.73/67.22 F1/EM score on the hidden test set
- The most-common mistake is answering un-answerable question. Adding a separate head for no-answer may help.