Our team combined Transformer-XL and QANet for the SQuAD QA system.

1. **Background**
   - Stanford Question Answering Dataset (SQuAD):
     - Reading comprehension dataset (Question Answering)
     - 100,000 Questions, over 50,000 unanswerable

2. **Methods**
   - **Baseline: Bidirectional Attention Flow (BiDAF)**
     - Embedding Layer: Projection & Highway
     - Encoder Layer: bidirectional LSTM
     - Attention Layer: bidirectional attention flow
     - Modeling Layer: bidirectional LSTM
     - Output Layer: bidirectional LSTM
   - **ConvNet character-level embedding**

ConvNet uses a sliding window to produce a feature map with each feature being
\[
   C_i = f(M - k_x + 1 + i)
\]
and is max-pooled to take the maximum value.

ConvNets are also used for word-level embeddings QANet and Transformer-XL.

(c) QANet
   - Transformer
   - Self-attention
   - Weight-sharing

(d) Transformer-XL

Key contributions:
- Segment-level recurrence mechanism
- Novel relative positional encoding scheme

\[
   b_i^{h_i} = \{h_i^{mcl}_i, \ldots, h_i^{mcl_1}\}
\]

\[
   q_i, k_i, v_i = \{h_i^{mcl_1}, \ldots, h_i^{mcl_1}\}, \{h_i^{mcl_1}, \ldots, h_i^{mcl_1}\}, \{h_i^{mcl_1}, \ldots, h_i^{mcl_1}\}
\]

\[
   \text{Attention} = \text{Q}V^T + \alpha \text{Q} \odot \text{K} + \beta \text{V}^T
\]

\[
   s_i^n = \text{Masked-Softmax}(A_i^n) v_i
\]

\[
   s_i^n = \text{LastNeuron}((v_i + h_i^{mcl_1}))
\]

\[
   h_i = \text{Position-wise Feed-Forward}(s_i^n)
\]

The segment-level recurrence shown above is added to QANet to realize the Transformer-XL model. The context is divided into segments by memory sequence lengths.

**Limitations for QA Tasks:**
- Segment context in QA tasks could break context continuity
- Previous segments have no information of later segments

3. **Quantitative Results**

4. **Prediction Error Analysis**
   - Insufficient understanding in sentence context
   - Ambiguities in context-question matching
   - Imprecise answer boundaries

**References**