# Gated Self-Matching Networks for Reading Comprehension and Question Answering

Presenter: Baosong Yang

# Motivation

- Focusing on reading comprehension style question answering which aims to answer questions given a passage or document.
- To account for the fact that words in the passage are of different importance: Gated Attention
- One answer candidate is often unaware of the clues in other parts of the passage: Self-Match

**Passage**: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. **When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901**, which he (Morgan) had caused. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

**Question**: On what did Tesla blame for the loss of the initial money?

**Answer**: Panic of 1901

# Architecture

- The recurrent network encoder (to build representation for questions and passage separately)
- The gated matching layer (to match the question and passage)

$$g_t = \text{sigmoid}(W_g[u_t^P, c_t])$$
$$[u_t^P, c_t]^* = g_t \odot [u_t^P, c_t]$$

- The self-matching layer (to aggregate information from the whole passage)
- The pointer network layer (to predict the start and end positions)

$$s_j^t = \mathrm{v}^{\mathrm{T}} \tanh(W_h^P h_j^P + W_{\mathrm{h}}^a h_{t-1}^a)$$
$$a_i^t = \exp(s_i^t) / \Sigma_{j=1}^n \exp(s_j^t)$$
$$p^t = \arg\max(a_1^t, \ldots, a_n^t)$$

# Architecture

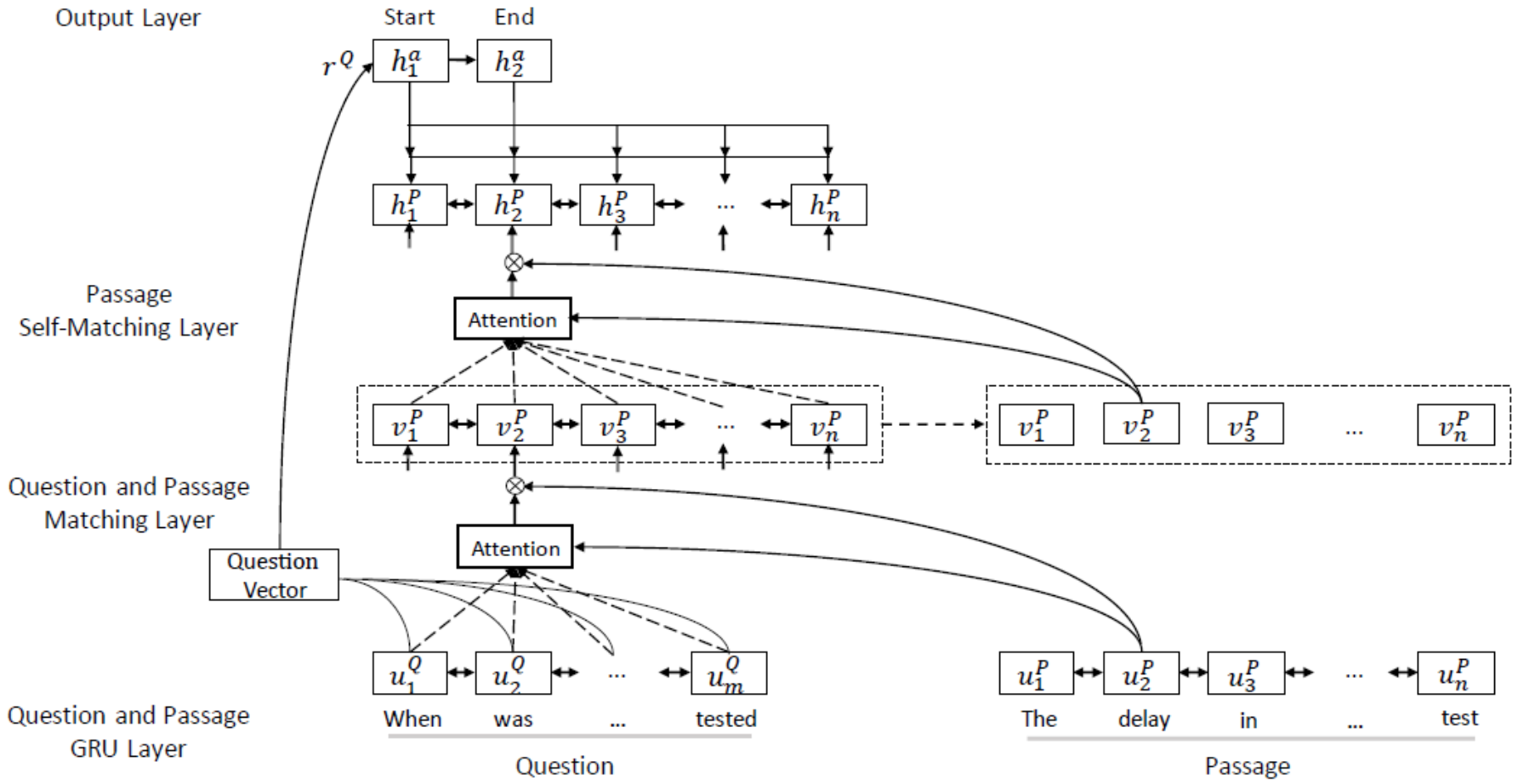



Figure 1: Gated Self-Matching Networks structure overview.

# Experiments

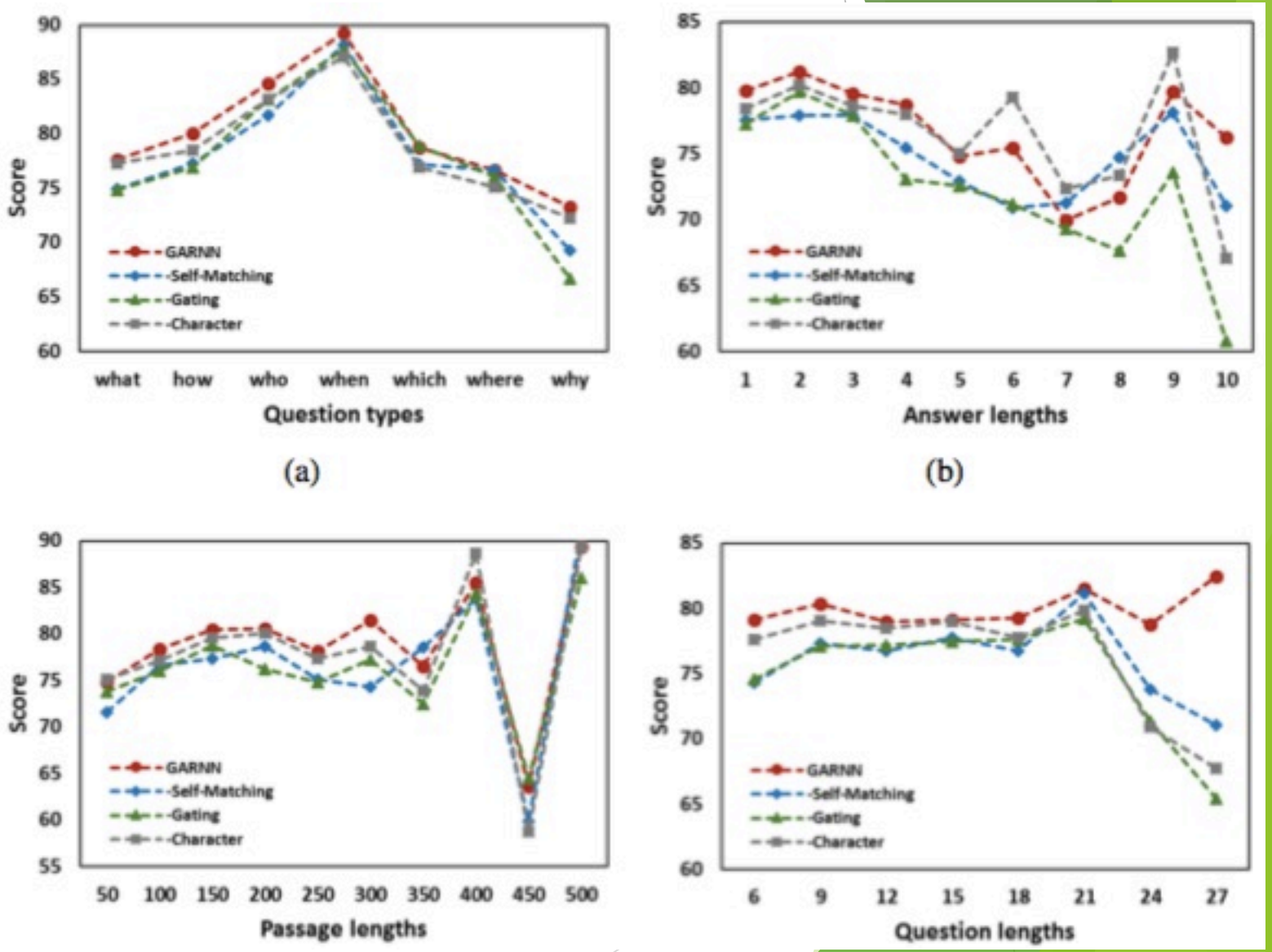
- SQuAD: 100,000+ questions (80%/10%/10%)

| | Dev Set | Test Set |
|---|---|---|
| *Single model* | **EM / F1** | **EM / F1** |
| LR Baseline (Rajpurkar et al., 2016) | 40.0 / 51.0 | 40.4 / 51.0 |
| Dynamic Chunk Reader (Yu et al., 2016) | 62.5 / 71.2 | 62.5 / 71.0 |
| Match-LSTM with Ans-Ptr (Wang and Jiang, 2016b) | 64.1 / 73.9 | 64.7 / 73.7 |
| Dynamic Coattention Networks (Xiong et al., 2016) | 65.4 / 75.6 | 66.2 / 75.9 |
| RaSoR (Lee et al., 2016) | 66.4 / 74.9 | - / - |
| BiDAF (Seo et al., 2016) | 68.0 / 77.3 | 68.0 / 77.3 |
| jNet (Zhang et al., 2017) | - / - | 68.7 / 77.4 |
| Multi-Perspective Matching (Wang et al., 2016) | - / - | 68.9 / 77.8 |
| FastQA (Weissenborn et al., 2017) | - / - | 68.4 / 77.1 |
| FastQAExt (Weissenborn et al., 2017) | - / - | 70.8 / 78.9 |
| **R-NET** | **71.1 / 79.5** | **71.3 / 79.7** |
| *Ensemble model* | | |
| Fine-Grained Gating (Yang et al., 2016) | 62.4 / 73.4 | 62.5 / 73.3 |
| Match-LSTM with Ans-Ptr (Wang and Jiang, 2016b) | 67.6 / 76.8 | 67.9 / 77.0 |
| RaSoR (Lee et al., 2016) | 68.2 / 76.7 | - / - |
| Dynamic Coattention Networks (Xiong et al., 2016) | 70.3 / 79.4 | 71.6 / 80.4 |
| BiDAF (Seo et al., 2016) | 73.3 / 81.1 | 73.3 / 81.1 |
| Multi-Perspective Matching (Wang et al., 2016) | - / - | 73.8 / 81.3 |
| **R-NET** | **75.6 / 82.8** | **75.9 / 82.9** |
| Human Performance (Rajpurkar et al., 2016) | 80.3 / 90.5 | 77.0 / 86.8 |

# Experiments

| Single Model | EM / F1 |
|---|---|
| **Gated Self-Matching (GRU)** | **71.1 / 79.5** |
| -Character embedding | 69.6 / 78.6 |
| -Gating | 67.9 / 77.1 |
| -Self-Matching | 67.6 / 76.7 |
| -Gating, -Self-Matching | 65.4 / 74.7 |

Table 3: Ablation tests of single model on the SQuAD dev set. All the components significantly (t-test, $p < 0.05$) improve the model.



# Conclusion

- Combine the idea of gate and attention.
- Speed and parameter size?
- A Gate layer after self-attention maybe also useful in NMT?