# Deep Co-attention Networks for Reading Comprehension

Amy Bearman[1]

## Reading Comprehension

**Problem**: Locate the answer to a question within a corresponding context paragraph

- **Training**: Using the Stanford Question Answering Dataset (SQuAD) dataset

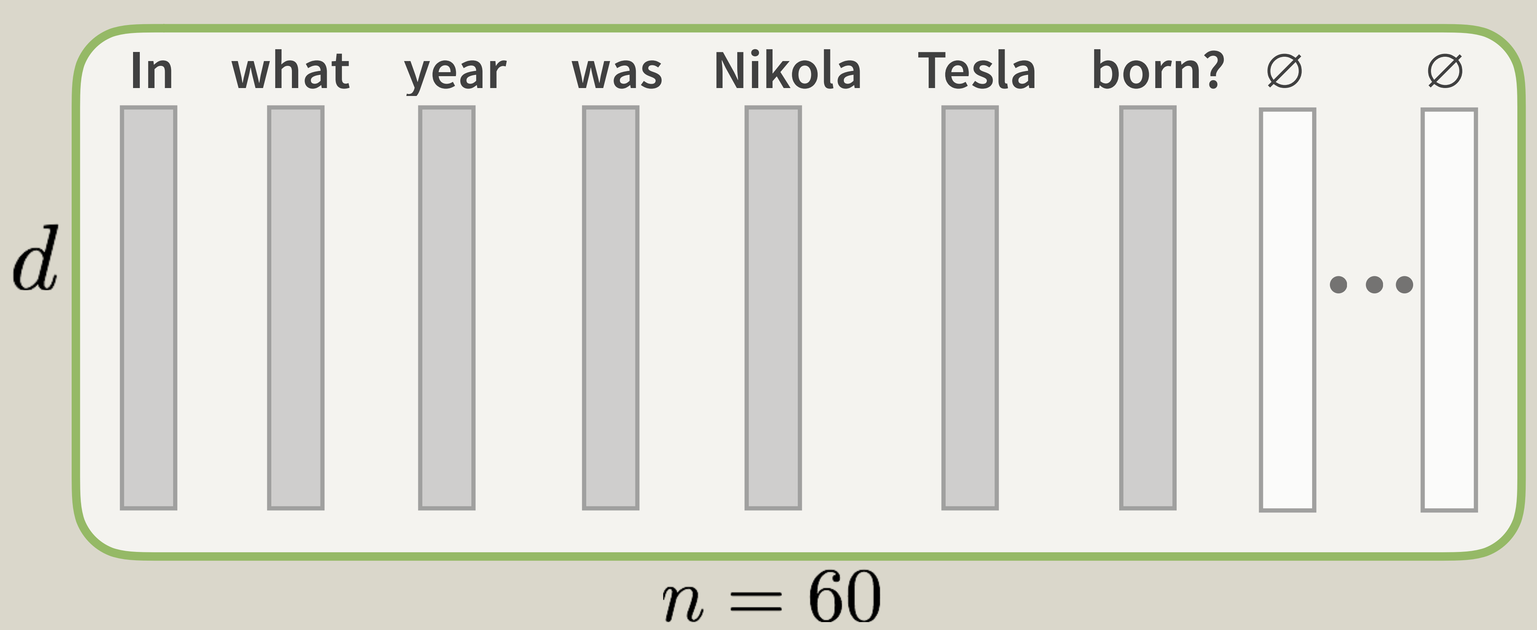
In what year was Nikola Tesla born?

Nikola Tesla (10 July **1856** - 7 January 1943) was a Serbian American inventor, electrical engineer, mechanical engineer, physicist, and futurist best known for his contributions to the design of the modern alternating current (AC) electricity supply system.

- **Output**: Predict the start and end index of the answer "span" in the paragraph: $a_s$ and $a_e$
- **Challenges**: The need for multi-sentence reasoning; maintaining long-range context, accounting for variable answer length

## Data Preprocessing

- **Tokenize** all questions, documents, and answers
- Obtain **GloVe word embeddings** for every word in the ~100K vocabulary, trimmed to dimension d = 100
- **Zero-pad** or truncate all questions to length n = 60 words, and all paragraphs to length m = 300




## Method

### Paragraph and Question Encoder

- **Question Encodings**: Encode the question word vectors using an LSTM, and stack the outputs horizontally.

$$q_t = \text{LSTM}_{enc}(q_{t-1}, x_t^Q) \qquad Q' = [q_1, q_2, \dots, q_n] \in \mathbb{R}^{\ell \times n}$$

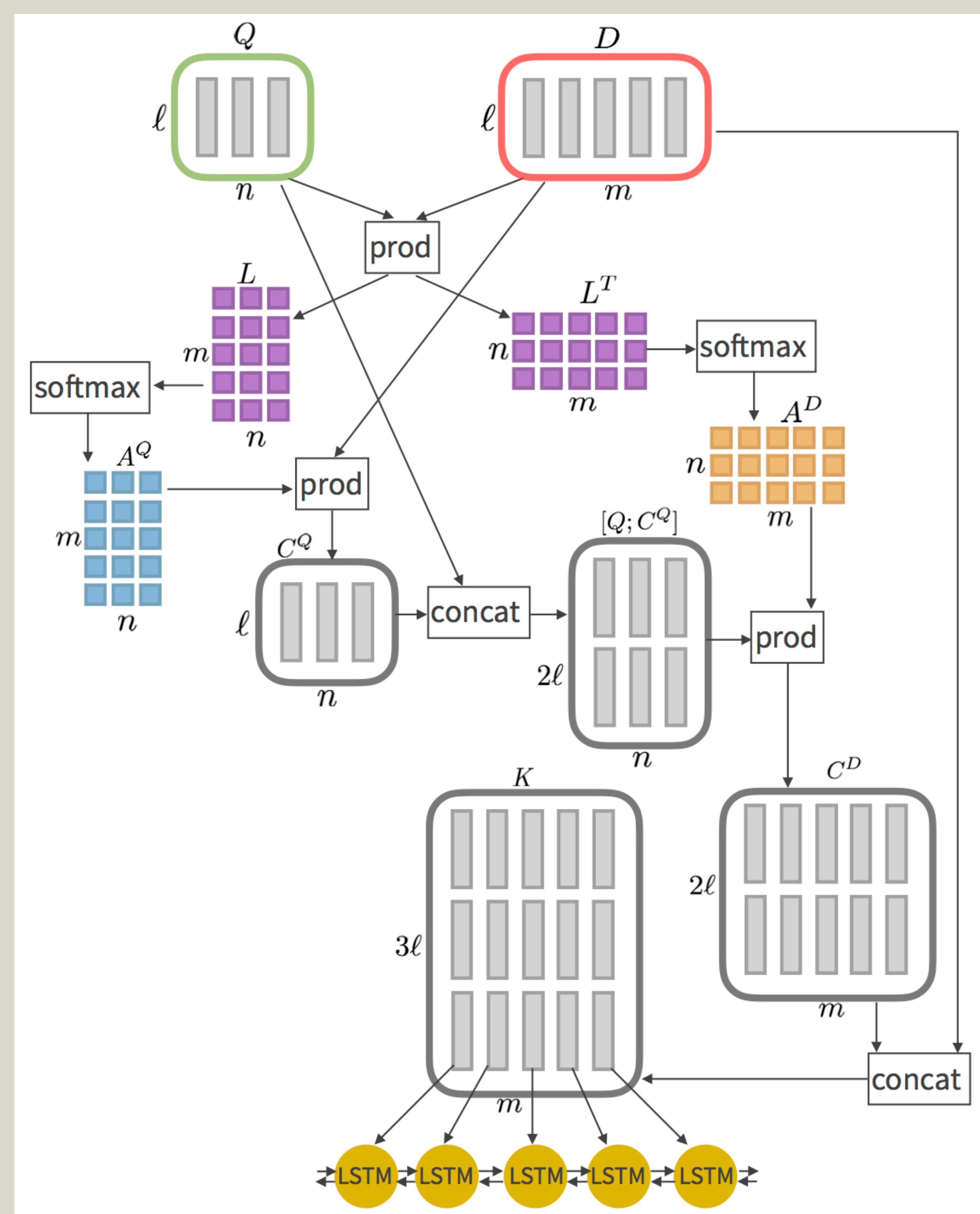
Add a nonlinear projection layer to the question encodings.

$$Q = \tanh(W^Q Q' + b^Q) \in \mathbb{R}^{\ell \times n}$$

- **Paragraph Encodings**: Encode the paragraph word vectors using the same LSTM, and stack the outputs horizontally.

$$d_t = \text{LSTM}_{enc}(d_{t-1}, x_t^D) \qquad D = [d_1, d_2, \dots, d_m] \in \mathbb{R}^{\ell \times m}$$

### Coattention Encoder




### Decoder

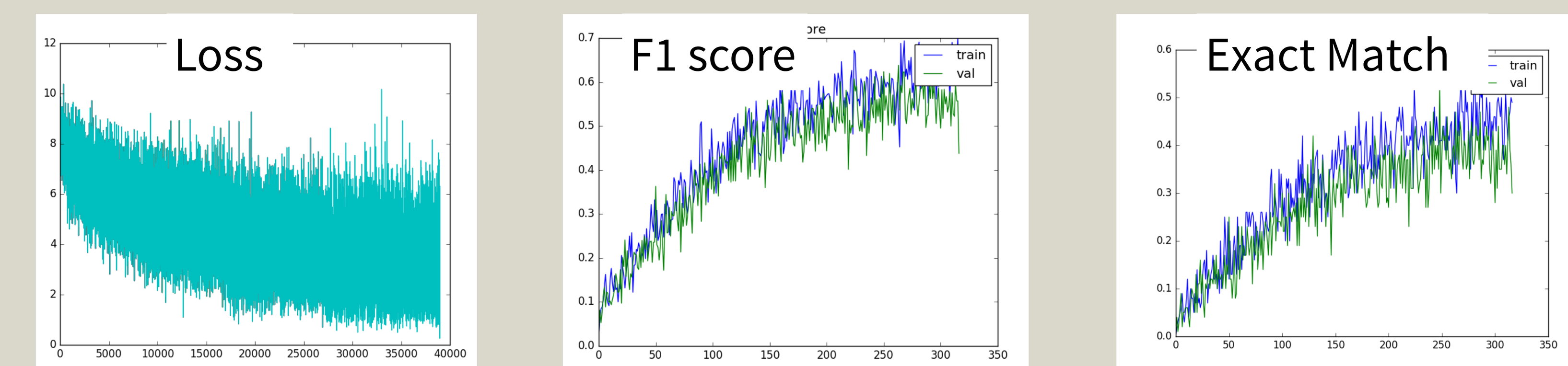
- Train two $m$-way classifiers: one to output $a_s$ and one for $a_e$

## Experimental Setup

- Optimize with Adam SGD for 10 epochs, with batch size = 20
- Hidden size = 200 for all LSTMs, and dropout rate = 0.1
- Initial learning rate of 1e-3, which is annealed over time
- Clip gradient norms at 10

## Results

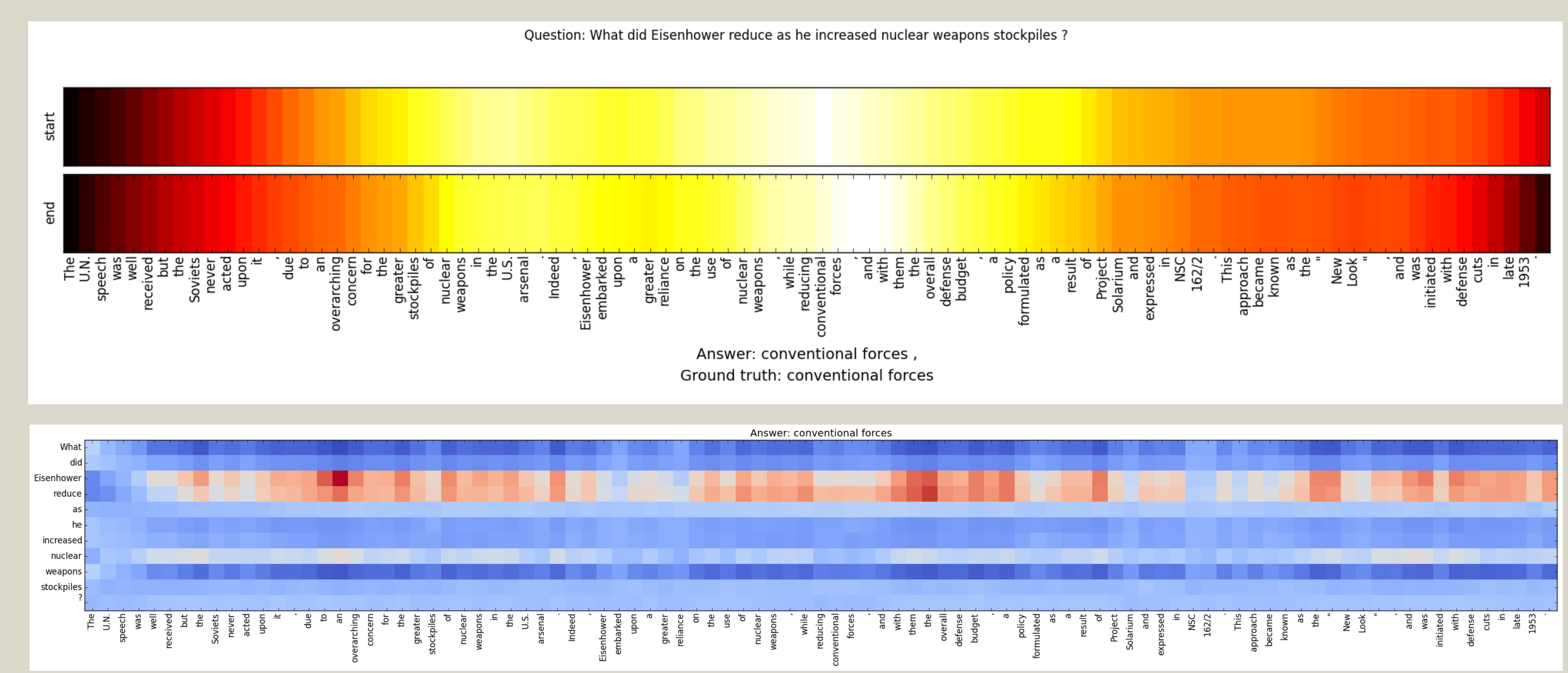
- **Metrics**: ExactMatch (% predictions that match GT exactly) and F1 Score (avg overlap between prediction and GT)




- **Quantitative Results**

| Model | Train F1 | Train EM | Dev F1 | Dev EM |
|---|---|---|---|---|
| Coattention, no dropout | 71.93 | 54.50 | 50.41 | 33.89 |
| Coattention, w/ dropout | | | **59.37** | **42.4** |

- **Visualizations**



## Future Work

- Add a more complicated decoder, so that the start and end answer are not in the same encoding space
- Add a simple constraint to ensure that $a_s \leq a_e$

abearman@cs.stanford.edu

[1]Stanford University