Lecture 11: Question Answering

Presenter: Chu Đình Đức

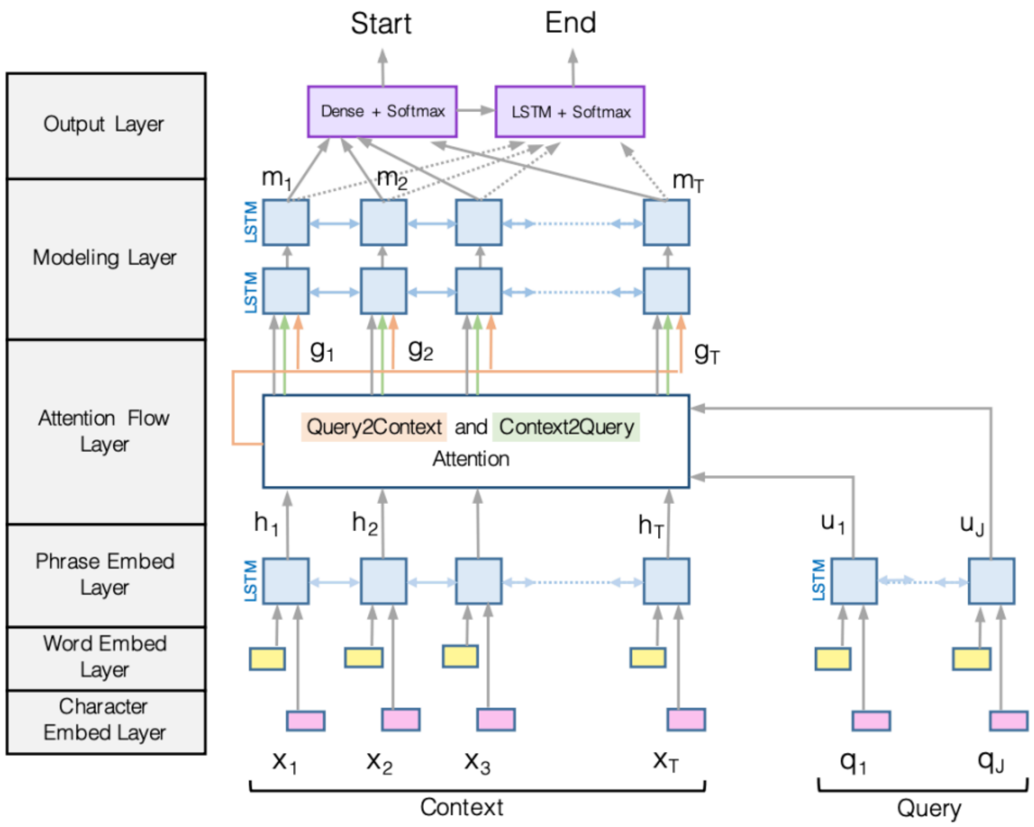
Lecture plan

1. BiDAF model
2. BERT model

Neural models for reading comprehension

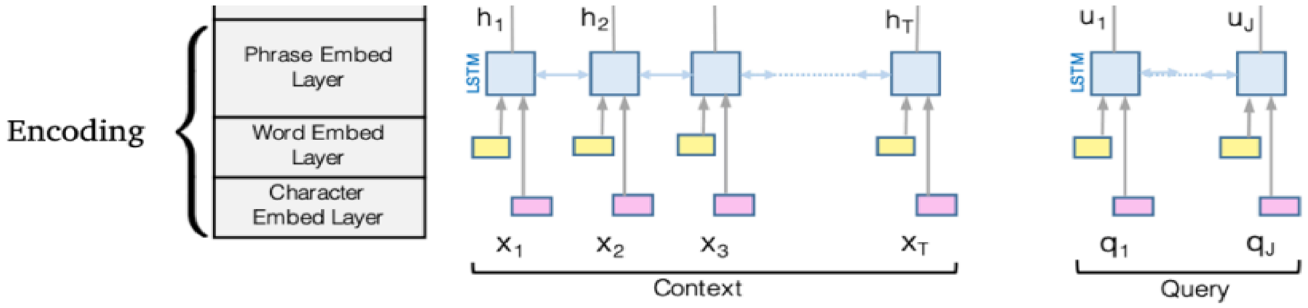
- Problem formulation
  - Input: $C = (c_1, c_2, \ldots, c_N)$ $Q = (q_1, q_2, \ldots, q_M)$, $c_i\ q_i \in V$
  - Output: $1 \leq start \leq end \leq N$
- 2 models
  - BiDAF (LSTM-based)
  - BERT

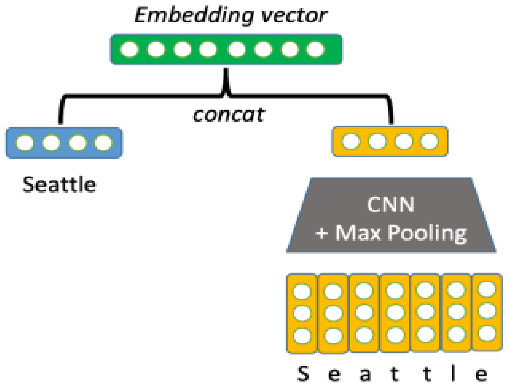
1. BiDAF: the Bidirectional Attention Flow model



(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

BiDAF: Encoding





- Use a concatenation of word embedding (GloVe) and character embedding (CNNs over character embeddings) for each word in context (C) and query (Q)

$$e(c_i) = f([GloVe(c_i);charEmb(c_i)]) \qquad e(q_i) = f([GloVe(q_i);charEmb(q_i)])$$

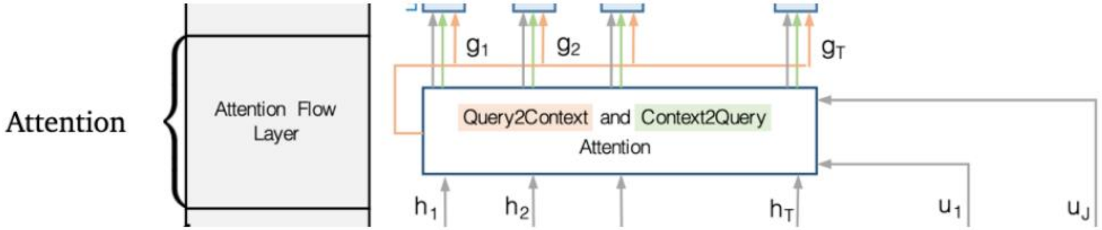
- Then, use two BiLSTMs separately to produce contextual embeddings for both context and query

$$\overrightarrow{\mathbf{c}}_i = \mathrm{LSTM}(\overrightarrow{\mathbf{c}}_{i-1}, e(c_i)) \in \mathbb{R}^H \qquad \overrightarrow{\mathbf{q}}_i = \mathrm{LSTM}(\overrightarrow{\mathbf{q}}_{i-1}, e(q_i)) \in \mathbb{R}^H$$

$$\overleftarrow{\mathbf{c}}_i = \mathrm{LSTM}(\overleftarrow{\mathbf{c}}_{i+1}, e(c_i)) \in \mathbb{R}^H \qquad \overleftarrow{\mathbf{q}}_i = \mathrm{LSTM}(\overleftarrow{\mathbf{q}}_{i+1}, e(q_i)) \in \mathbb{R}^H$$

$$\mathbf{c}_i = [\overrightarrow{\mathbf{c}}_i; \overleftarrow{\mathbf{c}}_i] \in \mathbb{R}^{2H} \qquad \mathbf{q}_i = [\overrightarrow{\mathbf{q}}_i; \overleftarrow{\mathbf{q}}_i] \in \mathbb{R}^{2H}$$

BiDAF: Attention



The final output is
$\mathbf{g}_i = [\mathbf{c}_i; \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{a}_i; \mathbf{c}_i \odot \mathbf{b}] \in \mathbb{R}^{8H}$

- First, compute a similarity score for every pair of $(\mathbf{c}_i, \mathbf{q}_j)$:

$$S_{i,j} = \mathbf{w}_{\text{sim}}^{\intercal}[\mathbf{c}_i; \mathbf{q}_j; \mathbf{c}_i \odot \mathbf{q}_j] \in \mathbb{R} \qquad \mathbf{w}_{\text{sim}} \in \mathbb{R}^{6H}$$

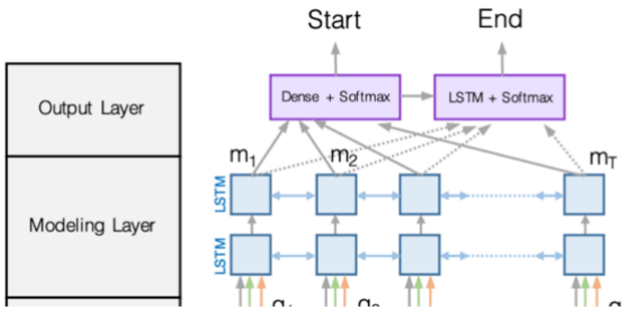
- Context-to-query attention (which question words are more relevant to c_i):

$$\alpha_{i,j} = \text{softmax}_j(S_{i,j}) \in \mathbb{R} \qquad \mathbf{a}_i = \sum_{j=1}^{M} \alpha_{i,j}\mathbf{q}_j \in \mathbb{R}^{2H}$$

- Query-to-context attention (which context words are relevant to some question words):

$$\beta_i = \text{softmax}_i(\max_{j=1}^{M}(S_{i,j})) \in \mathbb{R}^N \qquad \mathbf{b} = \sum_{i=1}^{N} \beta_i \mathbf{c}_i \in \mathbb{R}^{2H}$$

BiDAF: Modeling and output layers



The final training loss is

$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

Modeling layer: pass $\mathbf{g}_i$ to another two layers of **bi-directional** LSTMs.

- Attention layer is modeling interactions between query and context
- Modeling layer is modeling interactions within context words

$$\mathbf{m}_i = \text{BiLSTM}(\mathbf{g}_\text{i}) \in \mathbb{R}^{2H}$$

Output layer: two classifiers predicting the start and end positions:

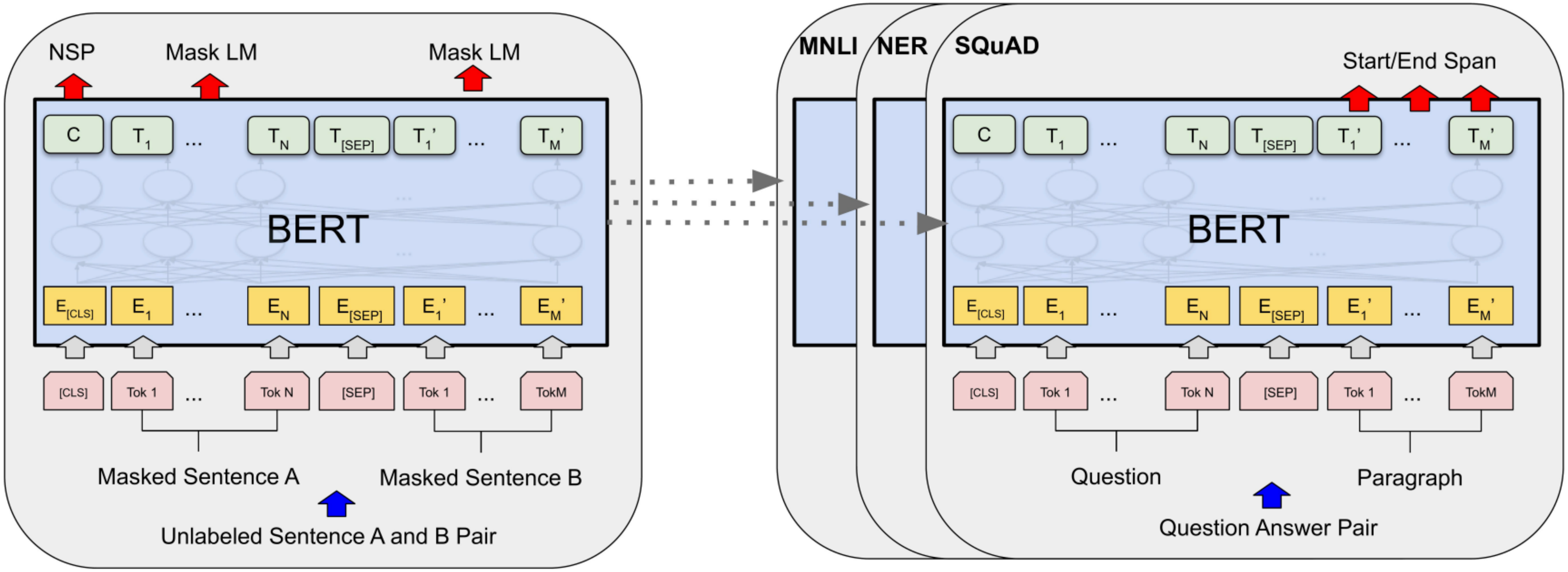
$$p_{\text{start}} = \text{softmax}(\mathbf{w}_{\text{start}}^\intercal[\mathbf{g}_i; \mathbf{m}_i]) \qquad p_{\text{end}} = \text{softmax}(\mathbf{w}_{\text{end}}^\intercal[\mathbf{g}_i; \mathbf{m}'_i])$$

$$\mathbf{m}'_i = \text{BiLSTM}(\mathbf{m}_i) \in \mathbb{R}^{2H} \quad \mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}} \in \mathbb{R}^{10H}$$

2. BERT

- BERT – Bidirectional Encoder Representations from Transformers
- BERT is a deep bidirectional Transformer encoder pre-trained on large amounts of text (Wikipedia + BooksCorpus)
- BERT is pre-trained on two training objectives:
  - Masked language model (MLM)
  - Next sentence prediction (NSP)
- $\text{BERT}_{\text{base}}$ has 12 layers and 110M parameters, $\text{BERT}_{\text{large}}$ has 24 layers and 330M parameters

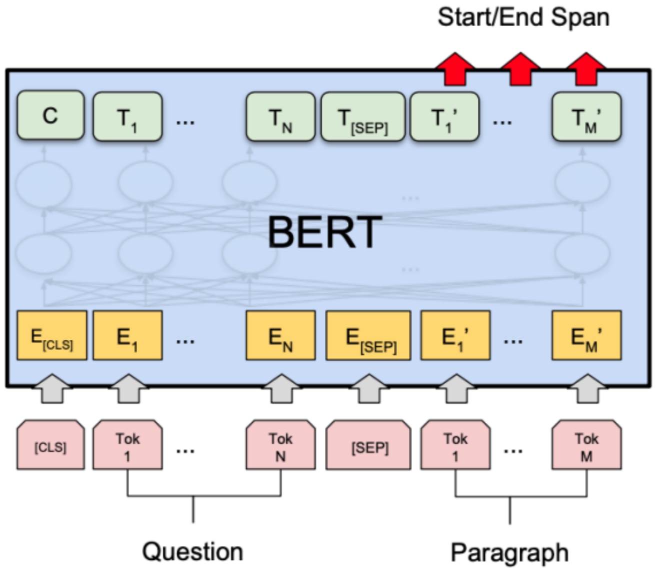
BERT



Pre-training

Fine-Tuning

BERT for reading comprehension



$$\mathcal{L} = -\log p_{\text{start}}(s^*) - \log p_{\text{end}}(e^*)$$

$$p_{\text{start}}(i) = \text{softmax}_i(\mathbf{w}_{\text{start}}^{\intercal}\mathbf{h}_i)$$

$$p_{\text{end}}(i) = \text{softmax}_i(\mathbf{w}_{\text{end}}^{\intercal}\mathbf{h}_i)$$

where $\mathbf{h}_i$ is the hidden vector of c_i, returned by BERT

Question = Segment A

Passage = Segment B

Answer = predicting two endpoints in segment B

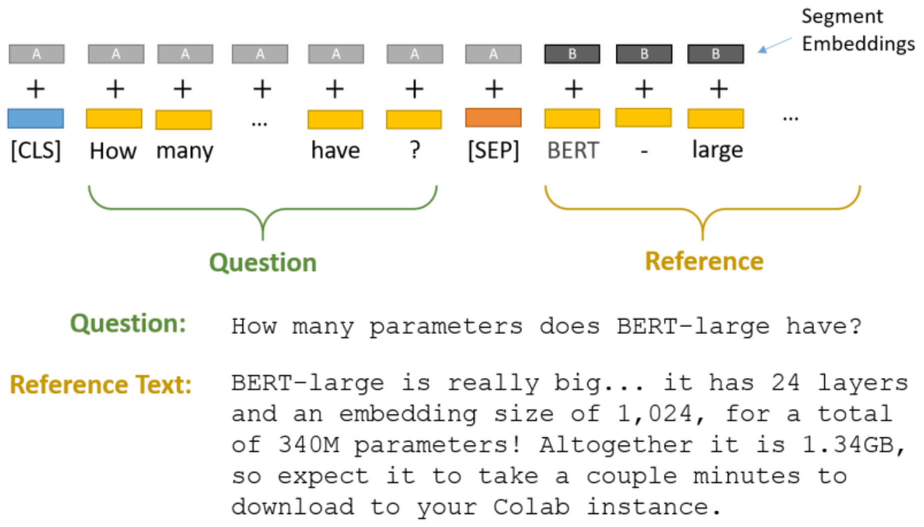
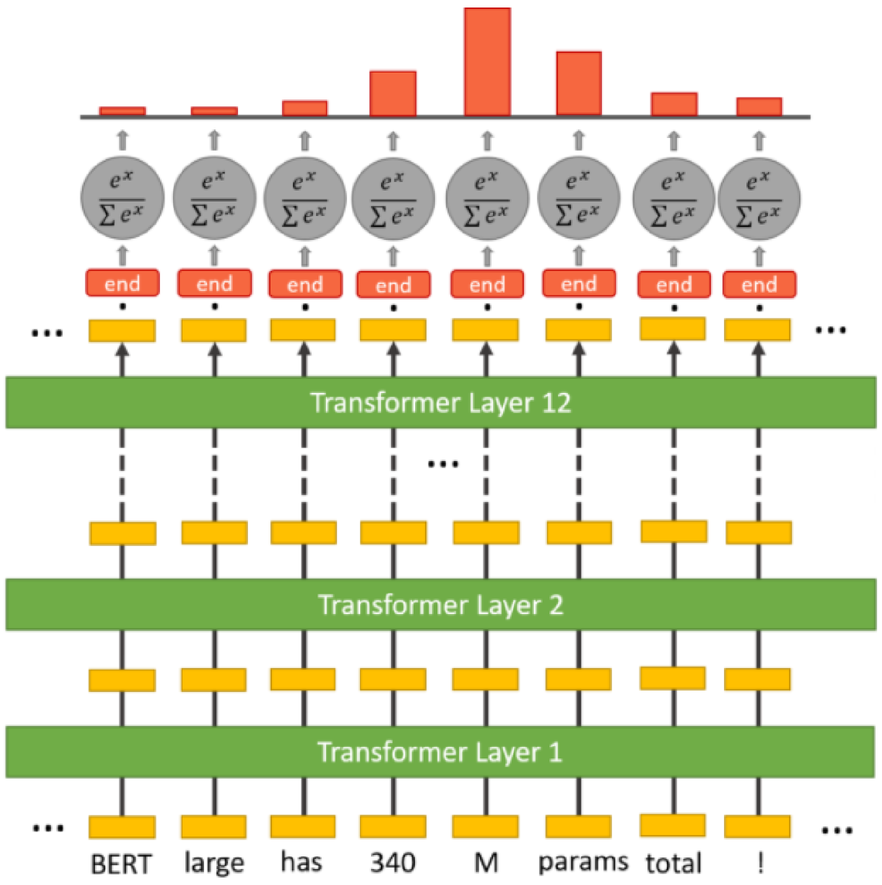


Image credit: https://mccormickml.com/

BERT for reading comprehension



THANK YOU

Any question?