input word indices. layers.embedding.

↓ embedding layer.

embeddings: $c_1, \cdots, c_N \in \mathbb{R}^D$ } V_i

$q_1, \cdots q_M \in \mathbb{R}^D$ }

↓ projection

$h_i = W_{proj} \cdot v_i \in \mathbb{R}^H$ ($W_{proj} \in \mathbb{R}^{H \times D}$, $v_i \in \mathbb{R}^D$)

↓ Highway Network x2

gate → $g = \sigma(W_g h_i + b_g) \in \mathbb{R}^H$ ($W_g \in \mathbb{R}^{H \times H}$)

transform → $t = \text{ReLu}(W_t h_i + b_t) \in \mathbb{R}^H$ ($W_t \in \mathbb{R}^{H \times H}$)

$h_i' = g \odot t + (1-g) \odot h_i \in \mathbb{R}^H$

↓ Encoder layer. layers.RNNEncoder.

bidirectional LSTM (1 layer)

$h_{i,fwd}' = \text{LSTM}(h_{i-1}', h_i) \in \mathbb{R}^H$

$h_{i,rev}' = \text{LSTM}(h_{i+1}', h_i) \in \mathbb{R}^H$

output: RNN's hidden state at each timestep i ← $h_i' = [h_{i,fwd}'; h_{i,rev}'] \in \mathbb{R}^{2H}$

↓ layers.BiDAFAttention

Attention Layer

$c_1, \cdots, c_N \in \mathbb{R}^{2H}$

$q_1, \cdots, q_M \in \mathbb{R}^{2H}$

$S_{ij} = w_{sim}^T [c_i; q_j; c_i \circ q_j] \in \mathbb{R}$ (element-wise product.)

($w_{sim} \in \mathbb{R}^{6H}$ weight vector)

C2Q:

$\bar{S}_{i,:} = \text{softmax}(S_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \cdots N\}$ (↳ attention distr.)

$a_i = \sum_{j=1}^{M} \bar{S}_{i,j} q_j \in \mathbb{R}^{2H} \quad \forall i \in \{1, \cdots, N\}$

↳ weighted sums of question hidden states. ↳ attention output

Q2C:

$\bar{\bar{S}}_{:,j} = \text{softmax}(S_{:,j}) \in \mathbb{R}^H \quad \forall j \in \{1, \cdots, M\}$

$S' = \bar{S} \cdot \bar{\bar{S}}^T \in \mathbb{R}^{N \times N}$

$b_i = \sum_{j=1}^{N} S_{ij}' c_j \in \mathbb{R}^{2H} \quad \forall j \in \{1, \cdots, M\}$

output ← $g_i = [c_i; a_i; c_i \cdot a_i; c_i \cdot b_i] \in \mathbb{R}^{8H} \quad \forall i \in \{1, \cdots, N\}$

↓ layers.RNNEncoder.

Modeling layer.

bidirectional LSTM (2 layer)

$m_{i,fwd} = \text{LSTM}(m_{i-1}, g_i) \in \mathbb{R}^H$

$m_{i,rev} = \text{LSTM}(m_{i+1}, g_i) \in \mathbb{R}^H$

$m_i = [m_{i,fwd}; m_{i,rev}] \in \mathbb{R}^{2H}$

↓ Layers.BiDAFOutput.

Output layer

bidirectional LSTM

$m_{i,fwd}' = \text{LSTM}(m_{i-1}', m_i) \in \mathbb{R}^H$

$m_{i,rev}' = \text{LSTM}(m_{i+1}', m_i) \in \mathbb{R}^H$

$m_i' = [m_{i,fwd}'; m_{i,rev}'] \in \mathbb{R}^{2H}$

↓ $G \in \mathbb{R}^{8H \times N}$ $[g_1 \cdots g_N]$

↓ $M, M' \in \mathbb{R}^{2H \times N}$ $[m_1 \cdots m_N]$

log-scale ← $P_{start} = \text{softmax}(W_{start}[G; M])$ ($W_{start} \in \mathbb{R}^{1 \times 10H}$)

Nll.lossfxn ← $P_{end} = \text{softmax}(W_{end}[G; M'])$