论文阅读笔记

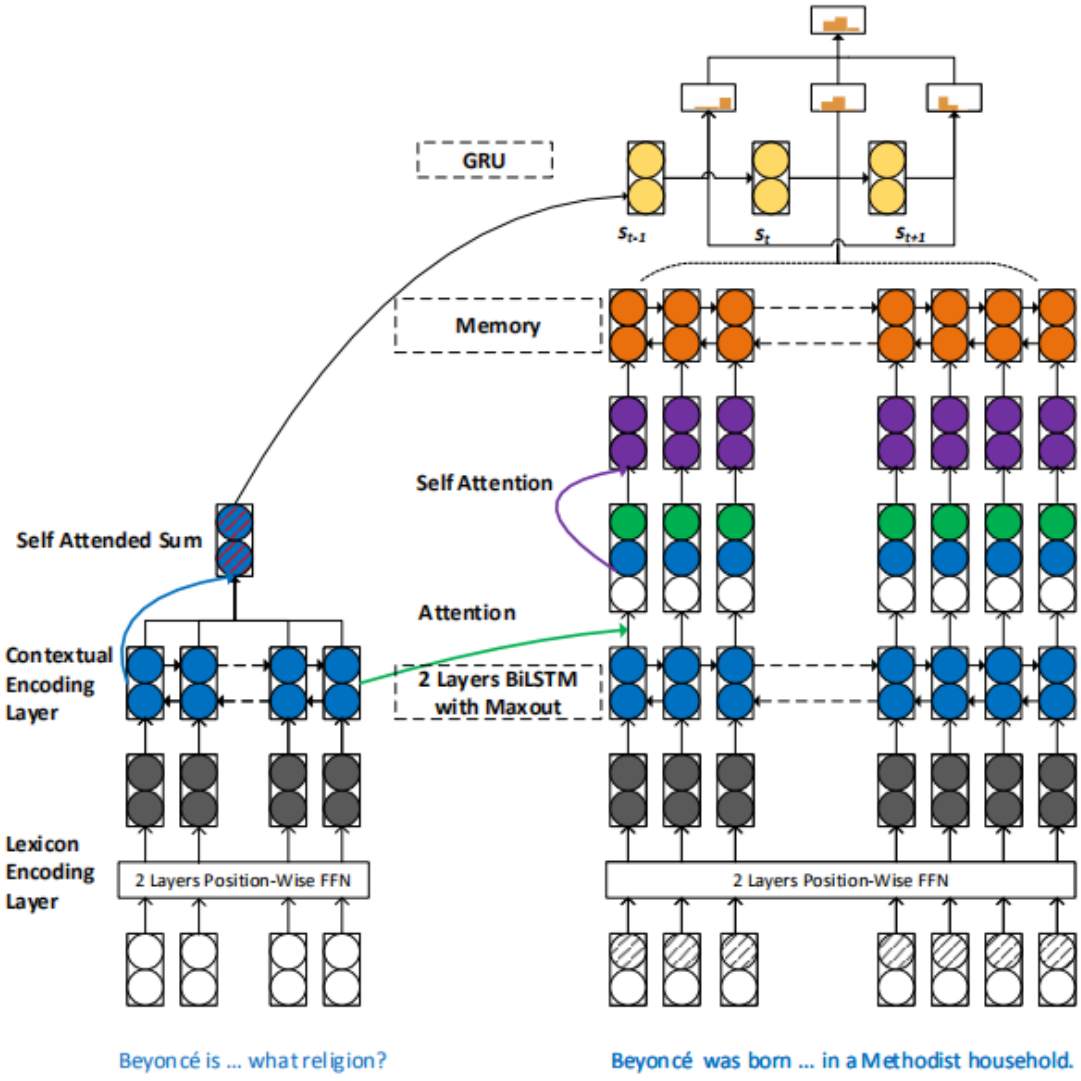
Step4

MF1833063, 史鹏, spwannasing@gmail.com

2019 年 6 月 30 日

1 Stochastic Answer Networks for Machine Reading Comprehension

相比以往使用了强化学习的multi-step推理，其独特的特点是在训练过程中，在神经网络的回答模块(最后一层)上使用了一种随机预测dropout。在训练过程中，我们确定了推理步骤的数目，但在答案模块上执行随机dropout，在解码过程中，我们根据所有步骤的预测平均值来生成答案，而不是最后一步。



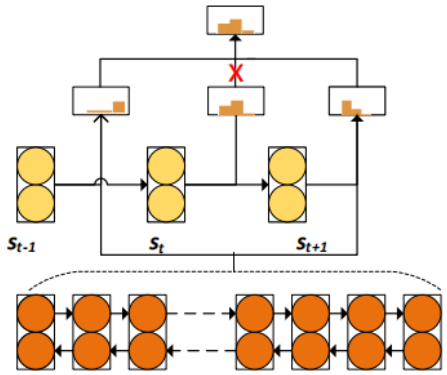


Figure 1: Illustration of "stochastic prediction dropout" in the answer module during training. At each reasoning step t, the model combines memory (bottom row) with hidden states $\mathbf{s_{t-1}}$ to generate a prediction (multinomial distribution). Here, there are three steps and three predictions, but one prediction is dropped and the final result is an average of the remaining distributions.

该算法的主要改进就在于answer prediction的过程中。

for $t \in \{0, 1, \ldots, T-1\}$

$$\begin{aligned} P_t^{begin} &= \text{softmax}\left(s_t W_6 M\right) \\ P_t^{end} &= \text{softmax}\left(\left[s_t; \textstyle\sum_j P_{t,j}^{begin} M_j\right] W_7 M\right) \end{aligned} \tag{1.1}$$

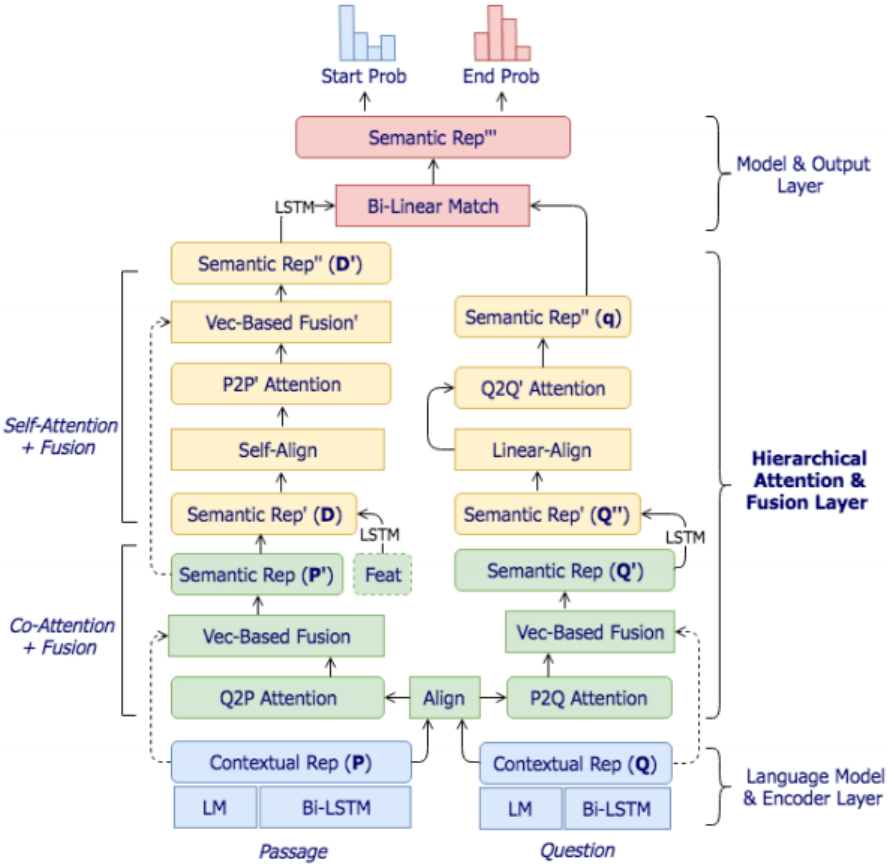
但是，与其从最后一步输出结果，我们使用所有的输出的平均得分。

$$\begin{aligned} P^{begin} &= \text{avg}\left(\left[P_0^{begin}, P_1^{begin}, \ldots, P_{T-1}^{begin}\right]\right) \\ P^{end} &= \text{avg}\left(\left[P_0^{end}, P_1^{end}, \ldots, P_{T-1}^{end}\right]\right) \end{aligned} \tag{1.2}$$

还有个特殊的是，在training的过程中，在averaging操作之前使用随机dropout，能够提升鲁棒性。

2 Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering

在所提出的方法中，注意和融合是水平和垂直地跨层进行的，在问题和段落之间的不同粒度级别上进行。具体来说，它首先用细粒度的语言嵌入对问题和段落进行编码，以便更好地捕捉语义级别上各自的表示。在此基础上，提出了一种多粒度融合方法，充分融合了全局表示和参与表示的信息。最后，介绍了一种分层关注网络，该网络以多层次的软度为中心，逐步解决了问题的答案跨度问题。



encoder-layer（这里的c_t可以看成是residual connection）

$$\begin{aligned} \mathrm{u}_{\mathrm{t}}^{\mathrm{Q}} &= \left[\mathrm{BiLSTMQ}\left(\left[\mathrm{e}_{\mathrm{t}}^{\mathrm{Q}}, \mathrm{c}_{\mathrm{t}}^{\mathrm{Q}}\right]\right), \mathrm{c}_{\mathrm{t}}^{\mathrm{Q}}\right] \\ \mathrm{u}_{\mathrm{t}}^{\mathrm{P}} &= \left[\mathrm{BiLSTM_P}\left(\left[\mathrm{e}_{\mathrm{t}}^{\mathrm{P}}, \mathrm{c}_{\mathrm{t}}^{\mathrm{P}}\right]\right), \mathrm{c}_{\mathrm{t}}^{\mathrm{P}}\right] \end{aligned} \tag{2.1}$$

Attention的意义：它的目的是使问题和段落对齐，以便我们能够更好地找到与该问题相关的最相关的passage span。

Co-attention&Fusion:

$$\mathrm{S_{ij}} = \mathrm{Att}\left(\mathrm{u_t^Q}, \mathrm{u_t^P}\right) = \mathrm{ReLU}\left(\mathrm{W_{lin}^\top u_t^Q}\right)^\top \cdot \mathrm{ReLU}\left(\mathrm{W_{lin}^\top u_t^P}\right) \tag{2.2}$$

P2Q Attention:

$$\alpha_\mathrm{j} = \mathrm{softmax}\left(\mathrm{S_{jj}}\right) \tag{2.3}$$

$$\tilde{\mathrm{Q}}_{:\mathrm{t}} = \sum_\mathrm{j} \alpha_{\mathrm{tj}} \cdot \mathrm{Q}_{:\mathrm{j}}, \forall \mathrm{j} \in [1, \ldots, \mathrm{m}] \tag{2.4}$$

Q2P Attention:

$$\begin{aligned} \beta_\mathrm{i} &= \mathrm{softmax}\left(\mathrm{S}_{\mathrm{i}:}\right) \\ \tilde{\mathrm{P}}_{\mathrm{k}:} &= \sum \beta_{\mathrm{ik}} \cdot \mathrm{P}_{\mathrm{i}:}, \forall \mathrm{i} \in [1, \ldots, \mathrm{n}] \end{aligned} \tag{2.5}$$

Fusion:

$$\begin{aligned} \mathrm{P}' &= \ \mathrm{Fuse}\ (\mathrm{P}, \tilde{\mathrm{Q}}) \\ \mathrm{Q}' &= \ \mathrm{Fuse}\ (\mathrm{Q}, \tilde{\mathrm{P}}) \end{aligned} \tag{2.6}$$

where Fuse $(\cdot, \cdot)$ is a typical fusion kernel

e.g.:

$$\mathrm{m}(\mathrm{P}, \tilde{\mathrm{Q}}) = \tanh\left(\mathrm{W_f}[\mathrm{P}; \tilde{\mathrm{Q}}; \mathrm{P} \circ \tilde{\mathrm{Q}}; \mathrm{P} - \tilde{\mathrm{Q}}] + \mathrm{b_f}\right) \tag{2.7}$$

Gating Function:

$$\begin{aligned} \mathrm{P}' &= \mathrm{g}(\mathrm{P}, \tilde{\mathrm{Q}}) \cdot \mathrm{m}(\mathrm{P}, \tilde{\mathrm{Q}}) + (1 - \mathrm{g}(\mathrm{P}, \tilde{\mathrm{Q}})) \cdot \mathrm{P} \\ \mathrm{Q}' &= \mathrm{g}(\mathrm{Q}, \tilde{\mathrm{P}}) \cdot \mathrm{m}(\mathrm{Q}, \tilde{\mathrm{P}}) + (1 - \mathrm{g}(\mathrm{Q}, \tilde{\mathrm{P}})) \cdot \mathrm{Q} \end{aligned} \tag{2.8}$$

Self-attention&Fusion:

$$\mathrm{D} = \mathrm{BiLSTM}\left([\mathrm{P}';\ \mathrm{feat\ man}\ ]\right) \tag{2.9}$$

$$\begin{aligned} \mathrm{L} &= \mathrm{softmax}\left(\mathrm{D} \cdot \mathrm{W}_1 \cdot \mathrm{D}^\top\right) \\ \tilde{\mathrm{D}} &= \mathrm{L} \cdot \mathrm{D} \end{aligned} \tag{2.10}$$

$$\mathrm{D}' = \mathrm{Fuse}(\mathrm{D}, \tilde{\mathrm{D}}) \tag{2.11}$$

$$\mathrm{D}'' = \mathrm{BiLSTM}\left(\mathrm{D}'\right) \tag{2.12}$$

$$\begin{aligned} \gamma &= \mathrm{softmax}\left(\mathbf{w}_\mathrm{q}^\top \cdot \mathrm{Q}''\right) \\ \mathbf{q} &= \sum \gamma_\mathrm{j} \cdot \mathrm{Q}''_\mathrm{j}, \forall \mathrm{j} \in [1, \ldots, \mathrm{m}] \end{aligned} \tag{2.13}$$

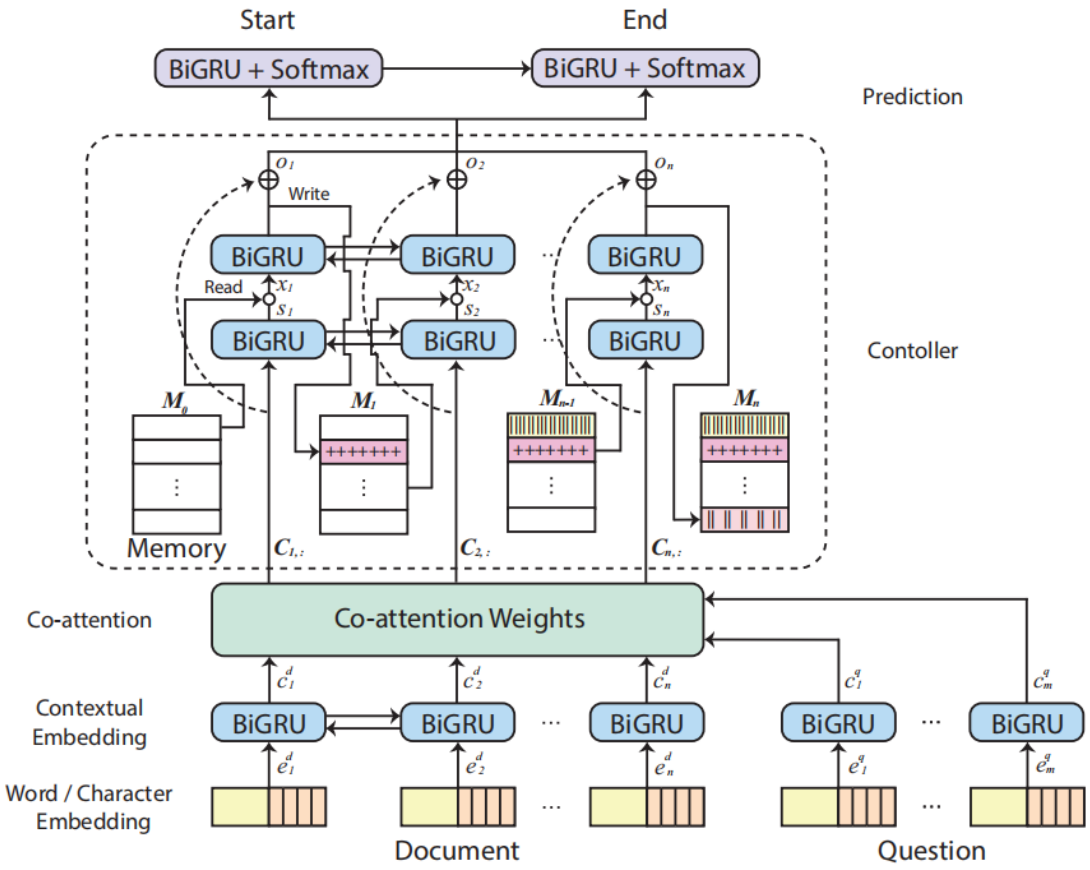
OutPut:

$$\begin{aligned} \mathrm{P}_{\ \mathrm{start}} &= \mathrm{softmax}\left(\mathbf{q} \cdot \mathbf{W}_\mathrm{s}^\top \cdot \mathrm{D}''\right) \\ \mathrm{P}_{\ \mathrm{end}} &= \mathrm{softmax}\left(\mathbf{q} \cdot \mathbf{W}_\mathrm{e}^\top \cdot \mathrm{D}''\right) \end{aligned} \tag{2.14}$$

$$\mathrm{L}(\theta) = -\frac{1}{\mathrm{N}} \sum_\mathrm{i}^\mathrm{N} \log \mathrm{p_s}\left(\mathrm{y_i^s}\right) + \log \mathrm{p_e}\left(\mathrm{y_i^e}\right) \tag{2.15}$$

3 A Multi-Stage Memory Augmented Neural Network for Machine Reading Comprehension

就是把MemoryNetwork换了一个领域来用。主要去学习Address、read、write的操作。



4 S-NET: FROM ANSWER EXTRACTION TO ANSWER GENERATION FOR MACHINE READING COMPREHENSION

提出了一个extraction-then-synthesis的框架。具体来说，答案提取模型首先用来预测文章中最重要的子区间作为证据，而答案综合模型则以证据作为附加的特征。并附上问题和段落，以进一步详细说明最后的答案。提出一项额外的passage-ranking任务（Multi-task learning的一次应用），以帮助在多个段落中提取答案。

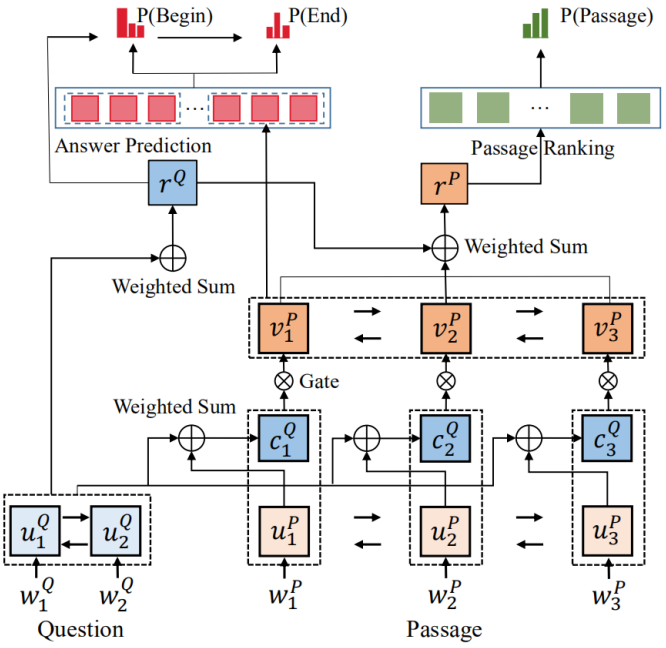


Figure 2: Evidence Extraction Model

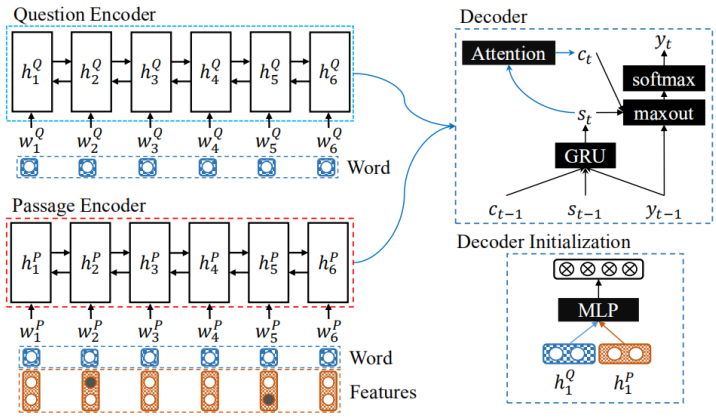


Figure 3: Answer Synthesis Model

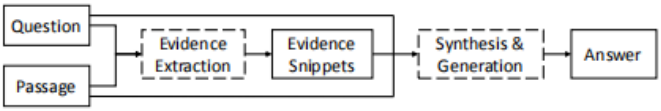


图 1: 整体结构

EVIDENCE SNIPPET PREDICTION：

$$\begin{aligned} u_t^Q &= \mathrm{BiGRU}_Q\left(u_{t-1}^Q, \left[e_t^Q, char_t^Q\right]\right) \\ u_t^P &= \mathrm{BiGRU}_P\left(u_{t-1}^P, \left[e_t^P, char_t^P\right]\right) \end{aligned} \tag{4.1}$$

attention-pooling vector:

$$\begin{aligned} s_j^t &= \mathrm{v}^{\mathrm{T}} \tanh\left(W_u^Q u_j^Q + W_u^P u_t^P\right) \\ a_i^t &= \exp\left(s_i^t\right) / \Sigma_{j=1}^m \exp\left(s_j^t\right) \\ c_t^Q &= \Sigma_{i=1}^m a_i^t u_i^Q \end{aligned} \tag{4.2}$$

sentence-pair representation:

$$v_t^P = \mathrm{GRU}\left(v_{t-1}^P, c_t^Q\right) \tag{4.3}$$

match-LSTM:

$$\begin{aligned} g_t &= \mathrm{sigmoid}\left(W_g\left[u_t^P, c_t^Q\right]\right) \\ \left[u_t^P, c_t^Q\right]^* &= g_t \odot \left[u_t^P, c_t^Q\right] \\ v_t^P &= \mathrm{GRU}\left(v_{t-1}^P, \left[u_t^P, c_t^Q\right]^*\right) \end{aligned} \tag{4.4}$$

Answer prediction:

$$\begin{aligned} s_j^t &= \mathrm{v}^{\mathrm{T}} \tanh\left(W_h^P v_j^P + W_h^a h_{t-1}^a\right) \\ a_i^t &= \exp\left(s_i^t\right) / \Sigma_{j=1}^N \exp\left(s_j^t\right) \\ p^t &= \mathrm{argmax}\left(a_1^t, \ldots, a_N^t\right) \end{aligned} \tag{4.5}$$

$$\begin{aligned} c_t &= \Sigma_{i=1}^N a_i^t v_i^P \\ h_t^a &= \mathrm{GRU}\left(h_{t-1}^a, c_t\right) \end{aligned} \tag{4.6}$$

$$\begin{aligned} s_j &= \mathbf{v}^{\mathrm{T}} \tanh\left(W_u^Q u_j^Q + W_v^Q v_r^Q\right) \\ a_i &= \exp\left(s_i\right) / \Sigma_{j=1}^m \exp\left(s_j\right) \\ r^Q &= \sum_{i=1}^m a_i u_i^Q \end{aligned} \tag{4.7}$$

$$\mathcal{L}_{AP} = -\Sigma_{t=1}^2 \Sigma_{i=1}^N \left[y_i^t \log a_i^t + \left(1 - y_i^t\right) \log\left(1 - a_i^t\right)\right] \tag{4.8}$$

Passage ranking:

$$\begin{aligned} s_j &= \mathrm{v}^{\mathrm{T}} \tanh\left(W_v^P v_j^P + W_v^Q r^Q\right) \\ a_i &= \exp\left(s_i\right) / \Sigma_{j=1}^n \exp\left(s_j\right) \\ r^P &= \Sigma_{i=1}^n a_i v_i^P \end{aligned} \tag{4.9}$$

$$g = v_g^{\mathrm{T}} \left(\tanh\left(W_g \left[r^Q, r^P\right]\right)\right) \tag{4.10}$$

$$\begin{aligned} \hat{g}_i &= \exp\left(g_i\right) / \Sigma_{j=1}^{k} \exp\left(g_j\right) \\ \mathcal{L}_{PR} &= -\sum_{i=1}^{k} \left[y_i \log \hat{g}_i + \left(1 - y_i\right) \log\left(1 - \hat{g}_i\right)\right] \end{aligned} \tag{4.11}$$

joint learning:

$$\mathcal{L}_E = r\mathcal{L}_{AP} + (1 - r)\mathcal{L}_{PR} \tag{4.12}$$

ANSWER SYNTHESIS:

$$\begin{aligned} h_t^P &= \mathrm{BiGRU}\left(h_{t-1}^P, [e_t^p, f_t^s, f_t^e]\right) \\ h_t^Q &= \mathrm{BiGRU}\left(h_{t-1}^Q, e_t^Q\right) \end{aligned} \tag{4.13}$$

$$\begin{aligned} s_j^t &= v_a^{\mathrm{T}} \tanh\left(W_a d_{t-1} + U_a h_j\right) \\ a_i^t &= \exp\left(s_i^t\right) / \Sigma_{j=1}^{n} \exp\left(s_j^t\right) \\ c_t &= \Sigma_{i=1}^{n} a_i^t h_i \end{aligned} \tag{4.14}$$

$$\begin{aligned} d_t &= \mathrm{GRU}\left(w_{t-1}, c_{t-1}, d_{t-1}\right) \\ d_0 &= \tanh\left(W_d \left[\overleftarrow{h}_1^P, \overleftarrow{h}_1^Q\right] + b\right) \end{aligned} \tag{4.15}$$

$$\begin{aligned} r_t &= W_r w_{t-1} + U_r c_t + V_r d_t \\ m_t &= \left[\max\left\{r_{t,2j-1}, r_{t,2j}\right\}\right]^{\mathrm{T}} \\ p\left(y_t | y_1, \ldots, y_{t-1}\right) &= \mathrm{softmax}\left(W_o m_t\right) \end{aligned} \tag{4.16}$$

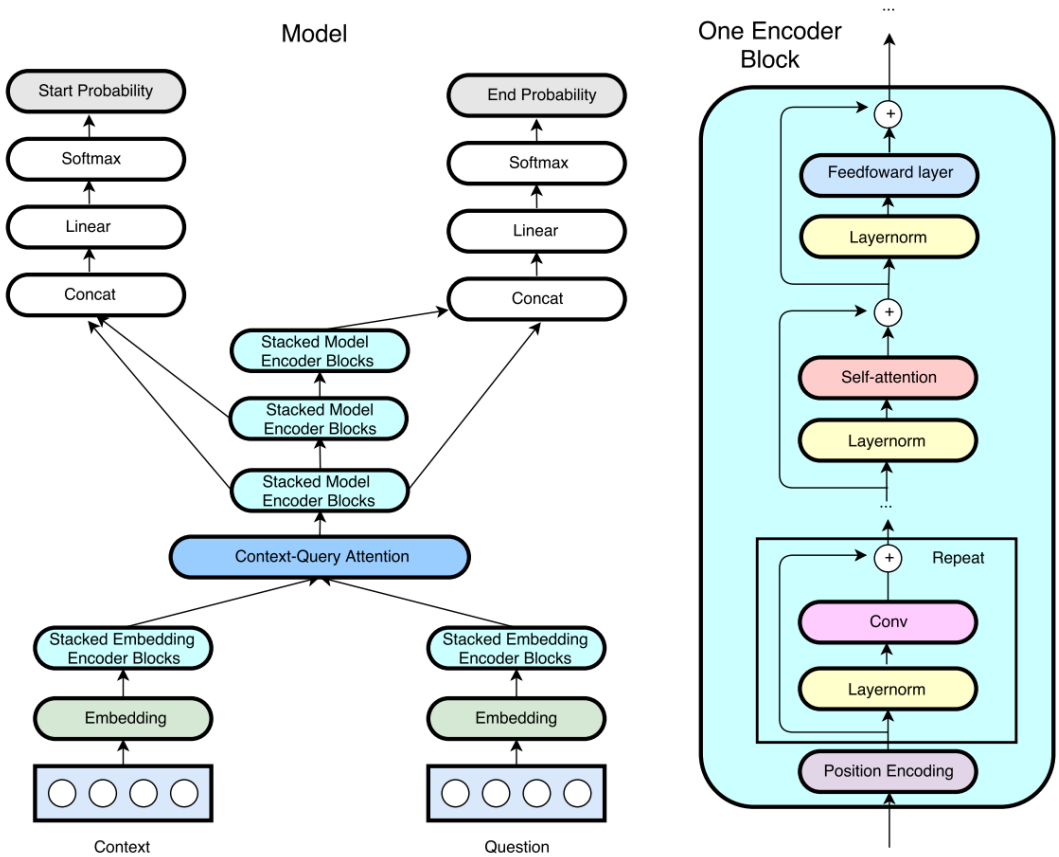
$$\mathcal{L}_S = -\frac{1}{|\mathcal{D}|} \Sigma_{(X,Y)\in\mathcal{D}} \log p(Y|X) \tag{4.17}$$

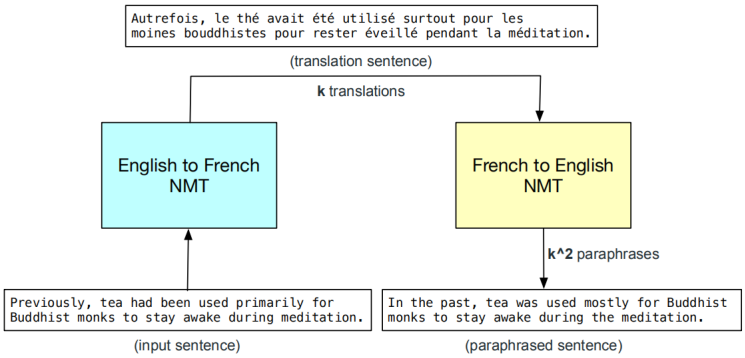
5 QANET: COMBINING LOCAL CONVOLUTION WITH GLOBAL SELF-ATTENTION FOR READING COMPREHENSION

本文的创新：

1）应该是首先在阅读理解任务上使用了类似Transformer的结构，抛弃了LSTM

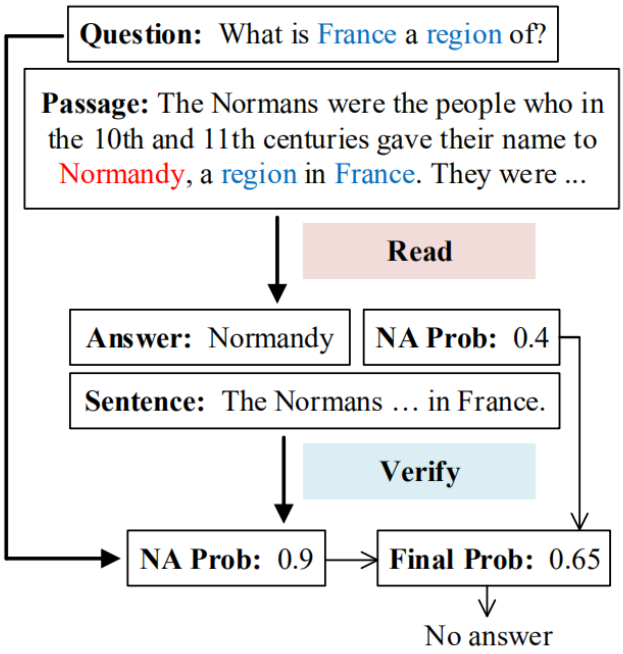
2）data augmentation，通过back translation来获得更多的训练数据。

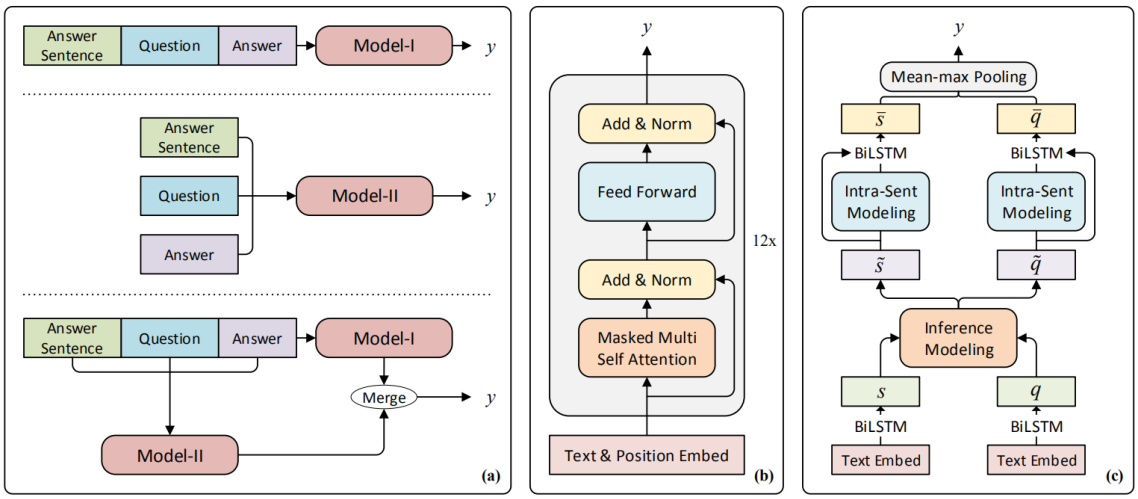




6 Read + Verify: Machine Reading Comprehension with Unanswerable Questions

机器阅读理解中有无法回答的问题，目的是在无法推断出答案时避免回答。然而，他们无法通过验证预测答案的合法性来验证问题的可回答性。本文的核心思想是提出了在原始的阅读理解模型的基础上，在预测出答案之后，再verify，重新计算一次No Ansewer的概率。





1. Reader with Auxiliary Losses

$$o_j = w_v^{\mathrm{T}} v_j, t = \sum_{j=1}^{l_q} \frac{e^{o_j}}{\sum_{k=1}^{l_q} e^{o_k}} v_j$$
$$\alpha, \beta = \text{ pointer network } (U, t) \tag{6.1}$$

$$\mathcal{L}_{\text{joint}} = -\log\left(\frac{(1-\delta)e^{z} + \delta e^{\alpha_a \beta_b}}{e^{z} + \sum_{i=1}^{l_p}\sum_{j=1}^{l_p} e^{\alpha_i \beta_j}}\right) \tag{6.2}$$

$$\tilde{o}_j = \tilde{w}_v^{\mathrm{T}} v_j, \tilde{t} = \sum_{j=1}^{l_q} \frac{e^{\tilde{o}_j}}{\sum_{k=1}^{l_q} e^{\tilde{o}_k}} v_j$$
$$\tilde{\alpha}, \tilde{\beta} = \text{ pointer network } (U, \tilde{t}) \tag{6.3}$$

$$\mathcal{L}_{\text{indep}-I} = -\log\left(\frac{e^{\tilde{\alpha}_{\overline{a}} \tilde{\beta}_{\overline{b}}}}{\sum_{i=1}^{l_p}\sum_{j=1}^{l_p} e^{\tilde{\alpha}_i \tilde{\beta}_j}}\right) \tag{6.4}$$

$$\mathcal{L}_{\text{indep}-II} = -(1-\delta)\log\sigma(z) - \delta\log(1-\sigma(z)) \tag{6.5}$$

$$\mathcal{L} = \mathcal{L}_{joint} + \gamma\mathcal{L}_{indep-I} + \lambda\mathcal{L}_{indep-II} \tag{6.6}$$

2. Answer Verifier

[1] Model-I: Sequential Architecture

将answer sentence，question，answer拼起来[S ; Q ; $; A]，用GPT的transformer（finetune）来执行此任务。

$$h_0 = W_e[X] + W_p$$
$$h_i = \text{ transformer\_block } (h_{i-1}), \forall i \in [1, n] \tag{6.7}$$

取出最后一层的最后一个token送入linear projection layer

$$p(y|X) = \text{softmax}\left(h_n^{l_m} W_y\right) \tag{6.8}$$

$$\mathcal{L}(\theta) = -\sum_{(X,y)} \log p(y|X) \tag{6.9}$$

[2] Model-II: Interactive Architecture we use a binary feature to indicate if a word is part of the answer(fea)

$$s_i = BiLSTM([word_i^s; char_i^s; fea_i^s]), \forall i \in [1, l_s]$$
$$q_i = BiLSTM([word_j^1; char_j^q; fea_j^q]), \forall j \in [1, l_q]$$

$$a_{ij} = s_i^{\top} q_j, \forall i \in [1, l_s], \forall j \in [1, l_q]$$
$$b_i = \sum_{j=1}^{l_q} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_q} e^{a_{ik}}} q_j, c_j = \sum_{i=1}^{l_s} \frac{e^{a_{ij}}}{\sum_{k=1}^{l_s} e^{a_{kj}}} s_i \tag{6.10}$$

$$\tilde{s}_i = F(s_i, b_i), \tilde{q}_j = F(q_j, c_j) \tag{6.11}$$

$$r = \text{gelu}\left(W_r[x; y; x \circ y; x - y]\right)$$
$$g = \sigma\left(W_g[x; y; x \circ y; x - y]\right) \tag{6.12}$$
$$o = g \circ r + (1 - g) \circ x$$

$$\overline{s}_i = \text{BiISTM}\left([\tilde{s}_i; \hat{s}_i]\right), \overline{q}_j = \text{BiLSTM}\left([\tilde{q}_j; \hat{q}_j]\right) \tag{6.13}$$

[3] Model-III: Hybrid Architecture

我们将两个模型的输出向量合并成一个单一的联合表示。然后应用统一的前馈分类器输出无答案概率.这样的设计让我们可以测试从两种不同架构的集成中可以获得更好的性能。在实践中，我们使用一个简单的连接来合并这两个信息源。

7 Adversarial Examples for Evaluating Reading Comprehension Systems

这篇论文主要就是验证了一个现象，当前的阅读理解模型对于原文添加了一些相关答案的对抗性句子(不违背原意)之后，准确率大幅下降。

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

8 Reading Wikipedia to Answer Open-Domain Questions

主要解决的是开放域问答，在这个论文里即：给出一个问题，从上万个wiki的document中找到答案，核心思想是分为两步，第一步document检索，找到和question最相关的几篇文章，然后第二步在这几篇document上运行常规的MRC算法即可。我们的方法将基于bigram散列和tf-idf匹配的搜索组件与经过训练以检测维基百科段落中的答案的多层递归神经网络模型相结合。

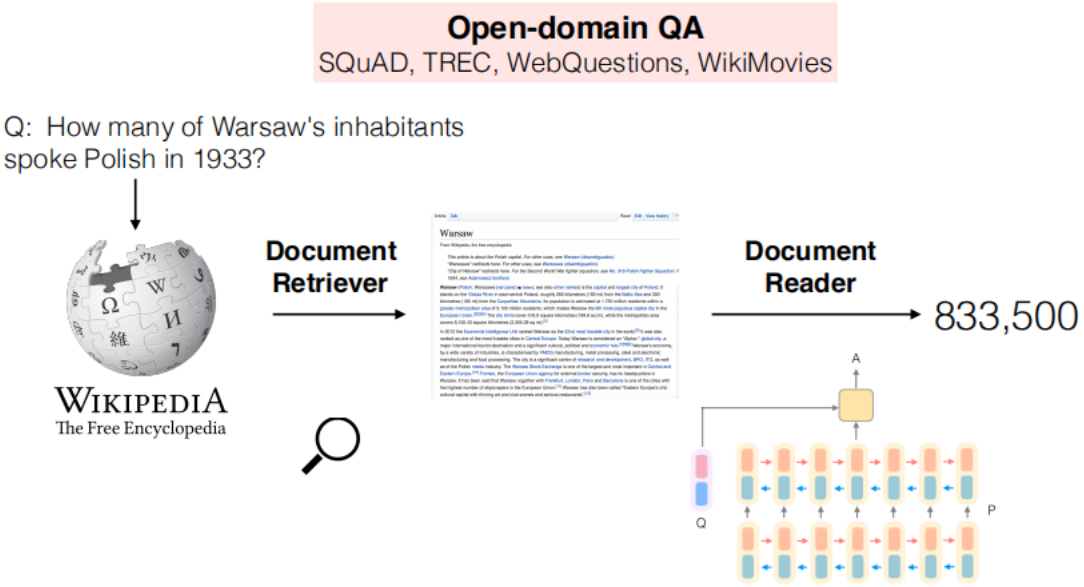
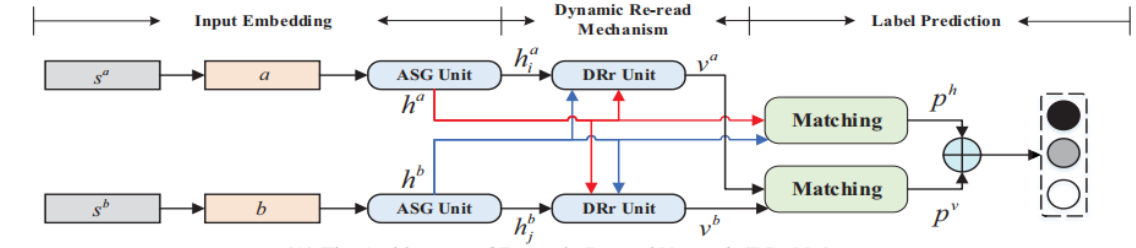


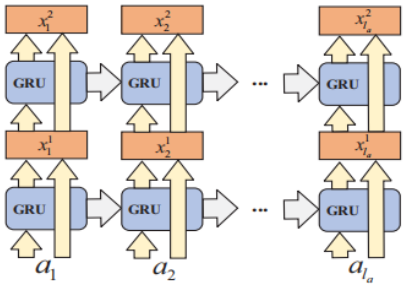
Figure 1: An overview of our question answering system DrQA.

9 DRr-Net: Dynamic Re-read Network for Sentence Semantic Matching

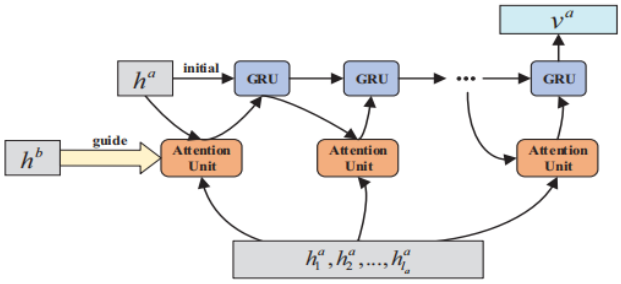
注意机制在捕捉语义关系和正确对齐两个句子的元素方面起着重要的作用。然而，句子在语义匹配过程中的重要部分随着句子理解程度的变化而动态变化。因此提出了DRr-Net，每一步都要注意句子的一个小区域，重读重要的单词，以便更好地理解句子的语义。重点是“Re”的思想，可以借鉴到MRC中。



(A) The Architecture of Dynamic Re-read Network (DRr-Net)



(B) Two-layer ASG Unit without attention



(C) DRr Unit

10 FLOWQA: GRASPING FLOW IN HISTORY FOR CONVERSATIONAL MACHINE COMPREHENSION

解决的问题的conversational阅读理解中对于之前的questions的记忆问题，通过“flow”来实现。核心思想是首先1个context对n个问题分别align，然后经过contextual（bi-lstm）处理后，不同question对应的contex的同一位置的单词送入lstm，即每个lstm处理同一个单词对不同question的align后的结果。重复此操作三次。

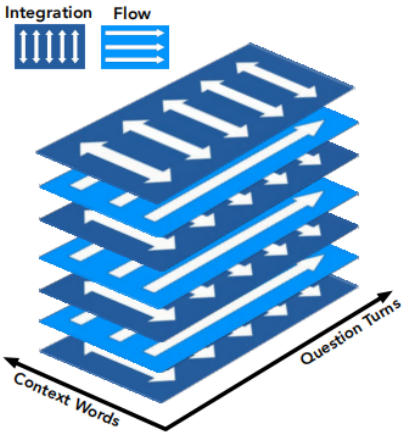


Figure 3: Alternating computational structure between context integration (RNN over context) and FLOW (RNN over question turns).

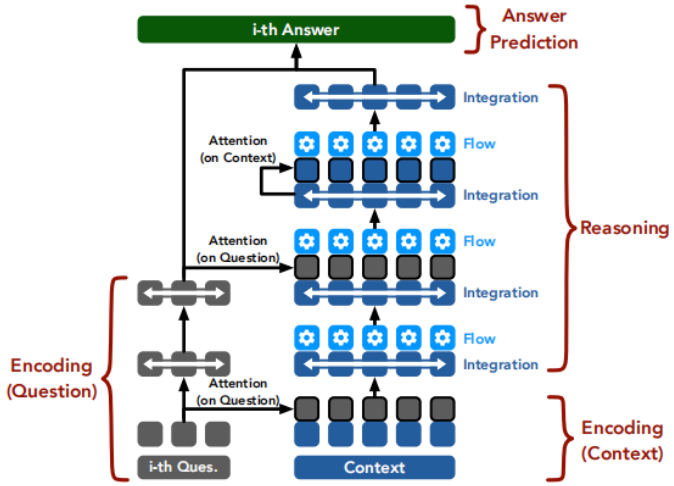


Figure 4: An illustration of the architecture for FLOWQA.

$$\hat{C}_i^h = \hat{c}_{i,1}^h, \ldots, \hat{c}_{i,m}^h = \text{BiLSTM}\left(\left[\boldsymbol{C}_i^h\right]\right) \tag{10.1}$$

$$f_{1,j}^{h+1}, \ldots, f_{t,j}^{h+1} = \text{GRU}\left(\hat{c}_{1,j}^h, \ldots, \hat{c}_{t,j}^h\right) \tag{10.2}$$

$$\begin{gathered} F_i^{h+1} = \left\{f_{i,1}^{h+1}, \ldots, f_{i,m}^{h+1}\right\} \\ C_i^{h+1} = c_{i,1}^{h+1}, \ldots, c_{i,m}^{h+1} = \left[\hat{c}_{i,1}^h; f_{i,1}^{h+1}\right], \ldots, \left[\hat{c}_{i,m}^h; f_{i,m}^{h+1}\right] \end{gathered} \tag{10.3}$$

$$Q_i^1 = q_{i,1}^1, \ldots, q_{i,n}^1 = \text{BiLSTM}\left(Q_i\right), Q_i^2 = q_{i,1}^2, \ldots, q_{i,n}^2 = \text{BiLSTM}\left(Q_i^1\right) \tag{10.4}$$

$$\tilde{q}_i = \sum_{k=1}^{n} \alpha_{i,k} \cdot q_{i,k}^2, \alpha_{i,k} \propto \exp\left(w^T q_{i,k}^2\right) \tag{10.5}$$

$$p_1, \ldots, p_t = \text{LSTM}\left(\tilde{q}_i, \ldots, \tilde{q}_t\right) \tag{10.6}$$

$$\begin{gathered} C_i^1 = \text{IF}\left(C_i^0\right) \\ C_i^2 = \text{IF}\left(C_i^1\right) \end{gathered} \tag{10.7}$$

$$\hat{q}_{i,j} = \sum_{k=1}^{n} \alpha^{i,j,k} \cdot q_{i,k}^2, \alpha^{i,j,k} \propto \exp\left(S\left(\left[c_i; c_{j,i}^1; c_{j,i}^2\right], \left[q_{j,k}; q_{j,k}^1; q_{j,k}^2\right]\right)\right) \tag{10.8}$$

$$C_i^3 = \text{IF}\left(\left[c_{i,1}^2; \hat{q}_{i,1}\right], \ldots, \left[c_{i,m}^2; \hat{q}_{i,m}\right]\right) \tag{10.9}$$

$$\hat{c}_{i,j} = \sum_{k=1}^{m} \alpha^{i,j,k} \cdot c_{i,k}^3, \alpha^{i,j,k} \propto \exp\left(S\left(\left[c_{i,j}^1; c_{i,j}^2, c_{i,j}^3\right], \left[c_{i,k}^1; c_{i,k}^2, c_{i,k}^3\right]\right)\right) \tag{10.10}$$

$$C_i^4 = \text{BiLSTM}\left(\left[c_{i,1}^3; \hat{c}_{i,1}\right], \ldots, \left[c_{i,m}^3; \hat{c}_{i,m}\right]\right) \tag{10.11}$$