Dynamic Coattention Networks For Question Answering – paper reproduction

D. Roy, R. Yeung, A. Poddar, K. Perlin, E. Chaumat
University of Oxford

***Abstract*— Hey how you doing. This is my abstract. Lorem ipsum.**

I. INTRODUCTION

Why did we select this paper? Paper summary: Set the experiment in context Underlines relevant technical details

II. IMPLEMENTATION

A. Dataset and Prepossessing

B. Model

Our Dynamic Coattention Network is divided into 5 units: the Encoder, the Coattention Module, the BiLSTM Encoder, the Highway Maxout Network, ant the Dynamic Pointer Decoder.

1) The Encoder: The questions $Q = (x_1^Q, x_2^Q, ..., x_n^Q) \in \Re^{B\times n}$ as well as their associated context documents $D = (x_1^D, x_2^D, ..., x_m^D) \in \Re^{B\times m}$ are taken as inputs, B corresponding to the batch size. Questions and documents are then encoded using the same LSTM :

$$d_t = LSTMenc(d_{t-1}, x_t^D) \in \Re^l \tag{1}$$

$$q_t = LSTMenc(q_{t-1}, x_t^Q) \in \Re^l \tag{2}$$

The encoded documents $D = [d_1, d_2, ..., d_m, d_\emptyset] \in \Re^{B\times l\times(m+1)}$ and $Q' = [q_1, q_2, ..., q_m, q_\emptyset] \in \Re^{B\times l\times(n+1)}$ are added $d_\emptyset$ and $q_\emptyset$ sentinels. The sentinels will allow to better predict rare and unseen words. The questions are then passed through non-linearities:$Q = tanh(W^{(Q)}Q' + b^{(Q)})$. This operation will permit some variation between the encoding space of the documents and the questions.

2) The Coattention Module: This module corresponds to the attention mechanism of the model. It permits to put focus for each document and each question on the relevant parts. The module takes the questions and the documents as inputs and then, outputs the co-attention context C^D, a shared representation of the attention of both the documents and questions.
First, L is computed:

$$L = D^T Q \in \Re^{B\times(m+1)\times(n+1)} \tag{3}$$

L represents the affinity score between each question and its associated document. L is then normalized row-wise to produce A^Q, the affinity weights for each word of the document with the words in the associated question. Likewise, A^D is obtained in normalizing L column-wise and represents for each word of the question, the affinity weights with the words of the associated document.

$$A^Q = softmax(L) \in \Re^{B\times(m+1)\times(n+1)} \tag{4}$$

$$A^D = softmax(L^T) \in \Re^{B\times(n+1)\times(m+1)} \tag{5}$$

These affinity coefficients represent the attention weights. Multiplied respectively with D and Q, the attention contexts for the documents and the questions are obtained.

$$C^Q = DA^Q \in \Re^{B\times l\times(n+1)} \tag{6}$$

Finally, the co-attention context $C^D \in \Re^{B\times 3l\times(m+1)}$ is computed by concatenating QA^D and C^QA^D.

3) The BiLSTM Encoder: This module will concatenate the documents and the co-attention context and encode them using a BiLSTM.

$$u_t = BiLSTM(u_{t-1}, u_{t+1}, [d_t; c_t^D]) \in \Re^{B\times 2l} \tag{7}$$

4) The Highway Maxout Network: This module takes the previous encoded start position $u_{s_{i-1}}$ and end position $u_{e_{i-1}}$, the co-attention encoding corresponding to the t^{th} word in the document u_t and the current hidden state of the Dynamic Pointer Decoder h_i and returns the start score α_t and end score β_t of the t^{th} word in the document.

$$\alpha_t = HMNstart(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}) \tag{8}$$

$$\beta_t = HMNend(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}) \tag{9}$$

The HMN is designed as follows:

$$HMN(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}) = \max(W^{(3)}[m_t^{(1)}; m_t^{(2)}] + b^{(3)}) \tag{10}$$

$$r = tanh(W^{(D)}[h_i; u_{s_{i-1}}; u_{e_{i-1}}]) \tag{11}$$

$$m_t^{(1)} = \max(W^{(1)}[u_t; r] + b^{(1)}) \tag{12}$$

$$m_t^{(2)} = \max(W^{(2)}m_t^{(1)} + b^{(2)}) \tag{13}$$

Add some more information about why this architecture is relevant here.

5) The Dynamic Pointer Decoder: The decoder takes as input the co-attention encoding U. At each iteration, the hidden state h_i is computed with the previous hidden state h_{i-1}, and the representation of the estimates of the start position $u_{s_{i-1}}$ and end position $u_{e_{i-1}}$.

$$h_i = LSTMdec(h_{i-1}, [u_{s_{i-1}}; u_{e_{i-1}}]) \tag{14}$$

Furthermore, the current start position and end position are computed using h_i, $u_{s_{i-1}}$, $u_{e_{i-1}}$ with the following equations:

$$s_i = \underset{t}{\text{argmax}}(\alpha_1, ..., \alpha_m) \tag{15}$$

$$e_i = \underset{t}{\text{argmax}}(\beta_1, ..., \beta_m) \tag{16}$$

C. Pipeline

D. Design decision

E. Difficulties and Remarks

III. Training and Evaluation

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

IV. Observations

How does it differ from the original paper? Why? What do we conclude?

V. Extensions

Lorem Ipsum is simply dummy text of the printing and typesetting industry. Lorem Ipsum has been the industry's standard dummy text ever since the 1500s, when an unknown printer took a galley of type and scrambled it to make a type specimen book. It has survived not only five centuries, but also the leap into electronic typesetting, remaining essentially unchanged. It was popularised in the 1960s with the release of Letraset sheets containing Lorem Ipsum passages, and more recently with desktop publishing software like Aldus PageMaker including versions of Lorem Ipsum.

VI. Conclusion

Remarks on the reproducibility the paper. Which parts of the paper can be reproduced, and at what cost in terms of resources (computation, time, people, development effort, communication with the authors). Now we're done, would we have done things differently?

References

[1] Xiong, Zhong, and Socher. "Dynamic Coattention Networks For Question Answering". ICLR. 2017.