MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE

MuBAF

Multi-Head Bi-Directional Attention Flow for Machine Reading Comprehension

Muhammad Umar Salman
umar.salman@mbzuai.ac.ae

Agenda

- Introduction
- Previous Work
- MuBAF
- Experiments
- Results
- Conclusion + Discussions

Introduction

- What is Question Answering?
- What is Machine Reading Comprehension?
- Span Prediction Problem
- Traditional
  - Syntactic Parsing, Pattern Matching, Question Classification
- State of the art:
  1. LUKE (Yamada et al.)
  2. SPAN-BERT (Joshi et al.)
  3. XLNET (Yang et al.)

Passage Sentence

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity.

Question

What causes precipitation to fall?

Answer Candidate

gravity

Problem + Motivation

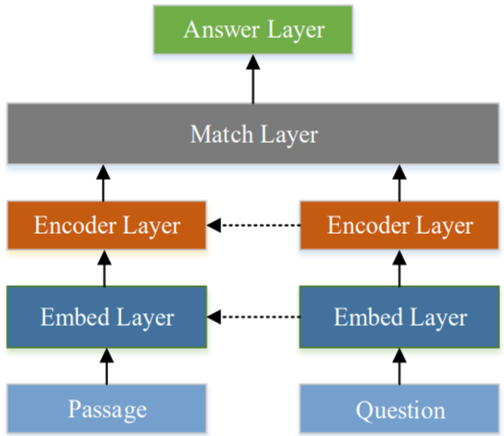
Why is BERT a problem?

- BERT is a huge model with over 100 million parameters
- Issues with QA in production
    1. GPU resources
    2. Thematic Structure, New Vocab and Writing Style
- BERT's pretraining are trained for general semantic
  (Not for domain specific terms e.g. Medicine and Law)
- Therefore we move back to LSTM and RNN architectures and build upon an existing architecture BiDAF

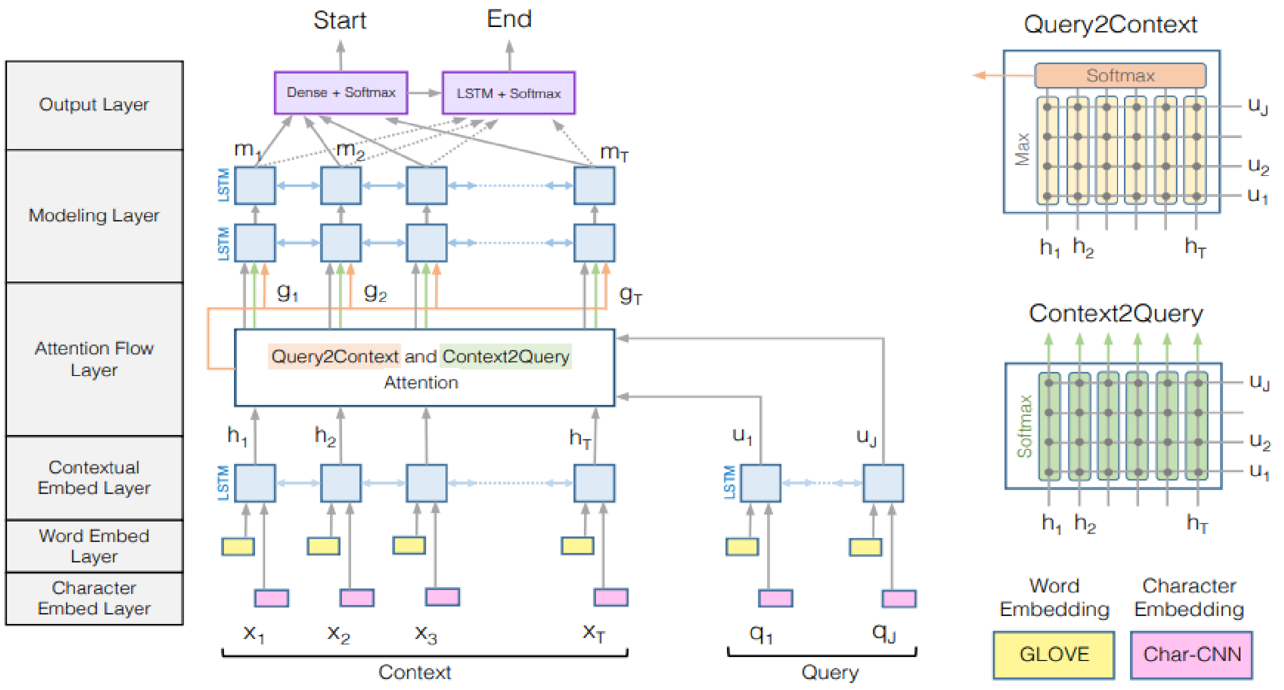
Previous Work

LSTM and RNN based MRC QA Systems

- Match LSTM Reader (Wang et al.)
- ReasonNet (Shen et al.)
- DCN+ (Xiong et al.)
- FusionNet (Huang et al.)
- BiDAF (Seo et al.)



BiDAF



Start
End
Dense + Softmax
LSTM + Softmax
Output Layer
Modeling Layer
Attention Flow Layer
Contextual Embed Layer
Word Embed Layer
Character Embed Layer
m1 m2 mT
LSTM
g1 g2 gT
Query2Context and Context2Query Attention
h1 h2 hT
u1 uJ
x1 x2 x3 xT
q1 qJ
Context
Query
Query2Context
Softmax
Max
Context2Query
Word Embedding
Character Embedding
GLOVE
Char-CNN

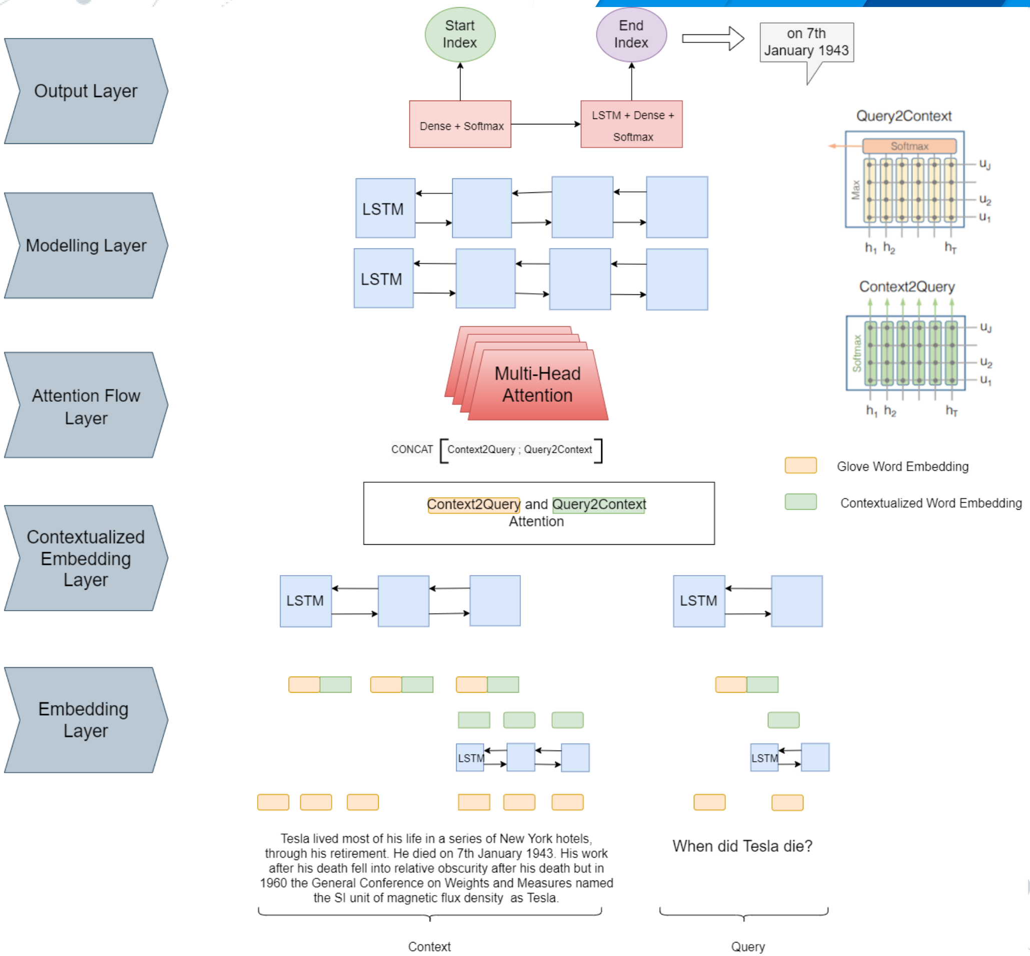
MuBAF

Motivation

To train effective and efficient QA system which can easily learn and adapt to change in writing style, theme as well as easily train on domain specific terms without the use of heavy GPU resources.

Contribution

- Use Contextualized Word Embedding's instead of Character Embeddings (Inspired by ELMo)
- Adopt a Multi Head Attention architecture instead of the already existing single attention layer (multiple representations)
- Make use of multi-layered fully connected layer to predict the indices of start and end of span.
- Created a cleaned and processed dataset with proper index labels, tokenization and POS and NER Tags .



Output Layer
Modelling Layer
Attention Flow Layer
Contextualized Embedding Layer
Embedding Layer
Start Index
End Index
on 7th January 1943
Dense + Softmax
LSTM + Dense + Softmax
LSTM
LSTM
Multi-Head Attention
CONCAT [Context2Query ; Query2Context]
Context2Query and Query2Context Attention
LSTM
LSTM
LSTM
LSTM
Tesla lived most of his life in a series of New York hotels, through his retirement. He died on 7th January 1943. His work after his death fell into relative obscurity after his death but in 1960 the General Conference on Weights and Measures named the SI unit of magnetic flux density as Tesla.
When did Tesla die?
Context
Query
Query2Context
Softmax
Max
Context2Query
Softmax
Glove Word Embedding
Contextualized Word Embedding

Experiments

1. SQuAD Dataset (Rajpurkar)
2. BiDAF PyTorch Implementation Starter Code
3. Flair Library for POS and NER Tags and Tokenization
4. GPU: Quadro RTX 6000

Training Dataset: 87599
Validation Dataset: 34726
Formatted Training: 86318
Formatted Validation: 34214

context	question	label	answer
Architecturally, the school has a Catholic cha...	To whom did the Virgin Mary allegedly appear i...	[515, 541]	Saint Bernadette Soubirous
Architecturally, the school has a Catholic cha...	What is in front of the Notre Dame Main Building?	[188, 213]	a copper statue of Christ
Architecturally, the school has a Catholic cha...	The Basilica of the Sacred heart at Notre Dame...	[279, 296]	the Main Building
Architecturally, the school has a Catholic cha...	What is the Grotto at Notre Dame?	[381, 420]	a Marian place of prayer and reflection
Architecturally, the school has a Catholic cha...	What sits on top of the Main Building at Notre...	[92, 126]	a golden statue of the Virgin Mary

Results

Reference	Model	Dev Set		Test Set	
		EM	F1	EM	F1
Wang et al. [13]	Match-LSTM	67.6	76.8	67.9	77.0
Seo et al. [11]	BiDAF	67.7	77.3	68.0	77.3
Shen et al. [16]	ReasoNet	70.8	79.4	69.1	78.9
Xiong et al. [17]	DCN+	74.5	83.1	75.1	83.1
Huang et al. [20]	FusionNet	75.3	83.6	76.0	83.9

Results

Model	Batch Size	Optimizer	# of Head	FC Layer	Contextual Embedding	EM	F1
Base	16	AdaDelta lr = 0.01	X	X	X	31.75	42.93
Run 1	16	AdaDelta	X	X	✓	55.21	60.44
Run 2	16	Adam	8	X	✓	42.25	46.63
Run 3	16	Adam	16	X	✓	38.79	42.37
Run 4	16	Adam	4	X	✓	43.81	47.34
Run 5	16	Adam	X	✓	✓	56.76	62.92
Run 6	8	Adam	X	✓	✓	54.54	59.32
Run 7	32	Adam	X	✓	✓	57.87	63.56

Conclusion + Discussion

Conclusion + Discussion

- From the results above we can see that the MHA didn’t us with the improvement we were expecting
- However, we can see using contextual embeddings inspired from what ELMo provides us with led to significant increase in the F1 and EM score.
- Adding a dense FC layer at the end of both the start index and the end index also increased the scores
- One reason why MHA failed to work was because we were applying it to the concatenation of C2Q and Q2C attention scores.
- Another reason for it not performing the way we had expected was because the MHA was not provided with a mask or positional encoding which we see in the transformer architecture

References

1. Shuohang Wang and Jing Jiang. Machine comprehension using match-lstm and answer pointer. arXiv preprint arXiv:1608.07905, 2016
2. Yelong Shen, Po-Sen Huang, Jianfeng Gao, and Weizhu Chen. Reasonet: Learning to stop reading in machine comprehension. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1047–1055, 2017.
3. Caiming Xiong, Victor Zhong, and Richard Socher. Dcn+: Mixed objective and deep residual coattention for question answering. arXiv preprint arXiv:1711.00106, 2017
4. Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, and Weizhu Chen. Fusionnet: Fusing via fully-aware attention with application to machine comprehension. arXiv preprint arXiv:1711.07341, 2017
5. Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603, 2016
6. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250, 2016

References

1. Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pretraining by representing and predicting spans.Transactions of the Association for Computational Linguistics, 8:64– 77, 2020
2. Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. Luke: deep contextualized entity representations with entity-aware self-attention. arXiv preprint arXiv:2010.01057, 2020
3. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32, 2019
4. Flair: https://github.com/flairNLP/flair
5. Starter: https://github.com/kushalj001/pytorch-question-answering/blob/master/2.%20BiDAF.ipynb

Thank you

Questions?

Muhammad Umar Salman
umar.salman@mbzaui.ac.ae