Who is Ernie and if so how many?

A multitasking Bert for question answering with discrete reasoning

Barthold Albrecht, Yanzhuo Wang, Xiaofang Zhu, {bholdia, yzw, zhuxf} @stanford.edu

Stanford
Computer Science

Predicting

With the introduction of pre-trained language models like BERT, the reading comprehension capabilities of algorithms has been significantly improved.
In this project, we are challenging a very new benchmark "DROP: Discrete Reasoning Over Paragraphs" which requires execute discrete operations like counting, sorting or comparing.
By adding 3 multi-task abilities to the BERT model, we improved the baseline from EM:24.2, F1:28.1 to EM:27.6, F1:30.7

Data

DROP provides dataset that consists of 96k question-answer pairs which have been adversarially created.
The dataset is crowdsourced, with answers labeled to all the questions.

References

[1]. Pranav Rajpurkar, et al, (2016). *Squad: 100,000+ questions for machine comprehension of text*. arXiv preprint arXiv:1606.05250
[2] Dua, D., Wang, Y., Dasigi, P., Stanovsky, G., Singh, S., Gardner, M, (2019).*DROP: A ReadingComprehension Benchmark Requiring Discrete Reasoning Over Paragraphs.* arXiv preprintarXiv:1903.00161

[3] Jacob Devlin, et al, (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805

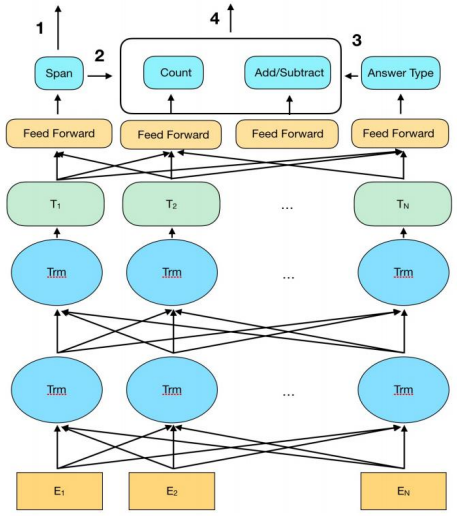
[4] Sebastian Ruder, (2017). *An Overview of Multi-Task Learning in Deep Neural Networks.* arXiv preprint arXiv:1903.00161

Features

There are 4 raw features from the input data: passage, question, answer and answer type. From these features, 3 more features are derived. For the answers that can be expressed as spans in the passages and questions, we derived the start/end indices for those spans. This is used by the span predictor. For examples with number or date type answers, we extract all the numbers from the passage, connect them by by “+” and “-” operators, compute all the expression combination, and record expressions that evaluate to the answer number. This is used by the task that predicts the arithmetic relationship between numbers in the passage.

Models

The embedded question and passage are encoded through the transformer blocks of BERT. The returned output sequence is then fed to four separate fully connected networks. If (1) the "Span" returns a prediction it is chosen as answer. If, however, (2)"Span" returns that the answer can not be found in the passage, an alternative answer will be chosen: the "Answer Type"-selector will decide whether the question can better be answered by "Count" or by "Add/Subtract" and (4) the answer is given accordingly.



Results

		Overall	Date (1.5%)	Number (61.9%)	Span (31.7%)	Spans (4.9%)
Baseline: BERT, SQuAD-style	EM	24.2	31.3	13.8	47.9	0
	F1	28.1	36.2	14.3	55.8	20.8
BERT + Count	EM	26.3	30.8	17.7	46.7	0
	F1	29.7	34.3	18	53.8	20.9
BERT + Count, + Q Span	EM	26.3	29.1	16.1	49.9	0
	F1	29.5	32.1	16.4	56.2	20.6
BERT + Count + Add/Sub	EM	27.6	29.9	17.8	50.7	0
	F1	30.7	31.2	18	56.8	21

Discussion

Overall, there is a moderate increase in both EM and F1 scores as we integrate more tasks to the BERT baseline model. This is expected.

1. the "Count" task substantially increases the model's performance on the "Number" type answers, by 3 to 4 points in both EM and F1 score.
2. adding the ability to predict answer as a question span has little effects on overall score.
3. The overall improvement from the Addition Subtraction task is also marginal.
4. Concerning the add/subtract capability of our model we experimented with different setups: for example feeding the extracted numbers of a passage along with the BERT output sequence to the add/subtract Feed Forward network. However, this did not to help much and seems to indicate that BERT in itself is sufficient to encode this relevant information directly.

Future

We believe that our model serves as a good platform for further improvements on the drop dataset. Therefore, we plan to further enhance the arithmetic capabilities of our model and to tackle the issue of multiple spans needed for the correct answer.