SQUAD 2.0
a fine tuning experiment

Cagan Bakirci
Minghao Liu
Jing Xiong

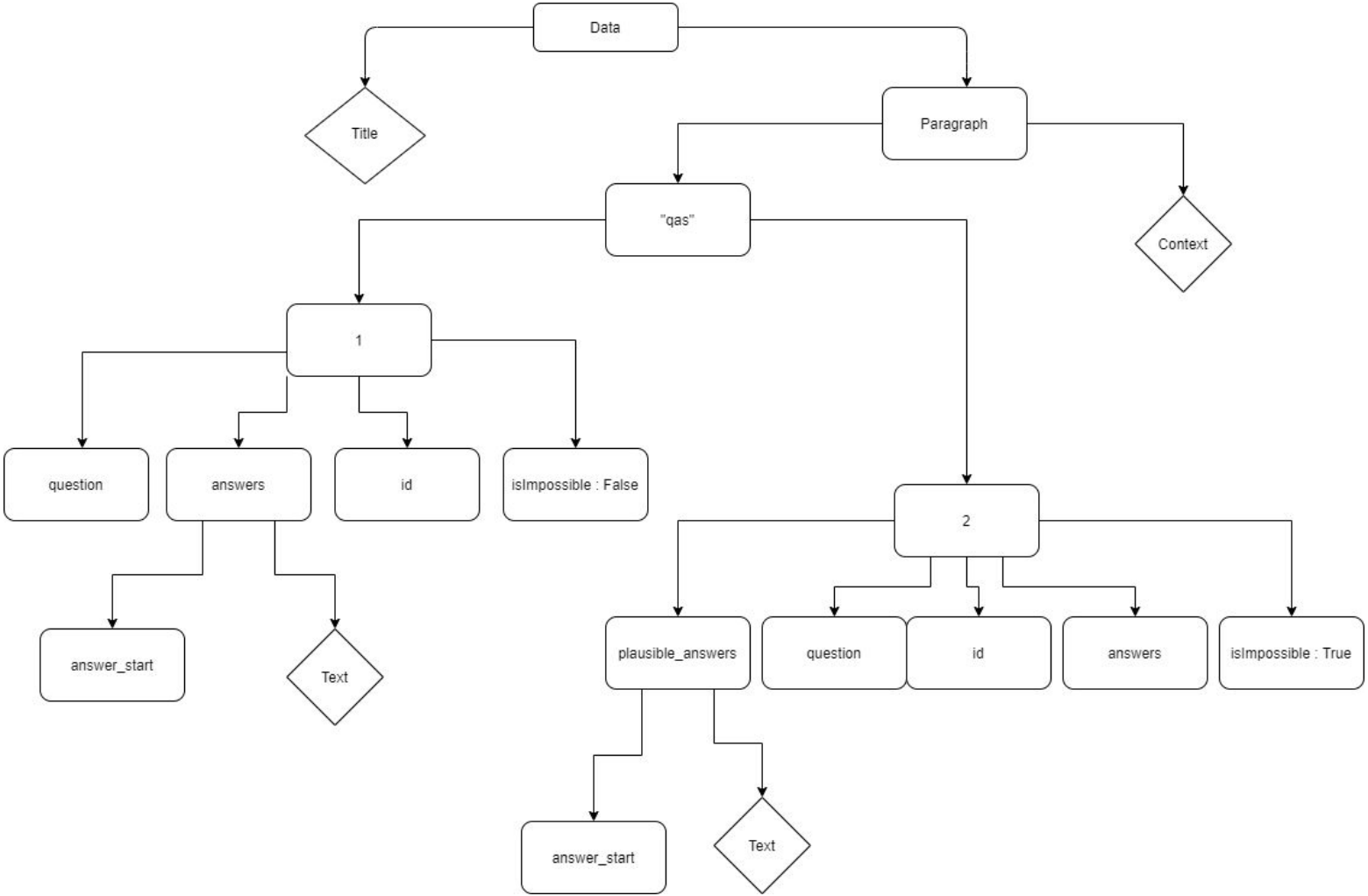
Dataset : SQUAD 2.0


```
{
    "qas": [
        {
            "question": "When was the Latin version of the word Norman first recorded?",
            "id": "56dde0379a695914005b9637",
            "answers": [
                {
                    "text": "9th century",
                    "answer_start": 309
                }
            ],
            "is_impossible": false
        },
        {
            "plausible_answers": [
                {
                    "text": "9th century",
                    "answer_start": 309
                }
            ],
            "question": "When was the French version of the word Norman first recorded?",
            "id": "5ad3ab70604f3c001a3feb8a",
            "answers": [],
            "is_impossible": true
        }
    ],
    "context": "The English name \"Normans\" comes from the French words Normans/Normanz,
    plural of Normant, modern French normand, which is itself borrowed from Old Low
    Franconian Nortmann \"Northman\" or directly from Old Norse Norðmaðr, Latinized
    variously as Nortmannus, Normannus, or Nordmannus (recorded in Medieval Latin, 9th
    century) to mean \"Norseman, Viking\"."
}
```


The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders


Data
Title
Paragraph
"qas"
Context
1
question
answers
id
isImpossible : False
answer_start
Text
2
plausible_answers
question
id
answers
isImpossible : True
answer_start
Text

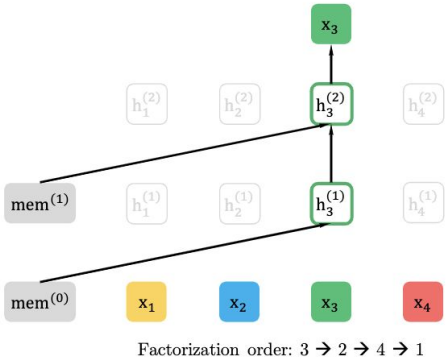
XLNet: Permutation language model

Permutation language model (AR +AE) instead of AE only(Bert)

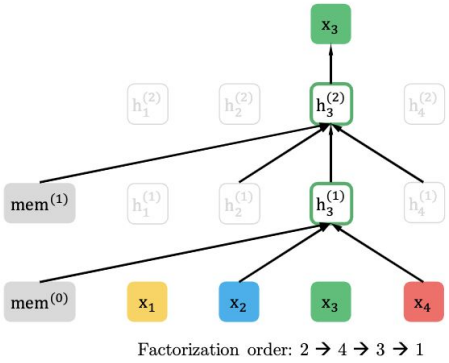
AR: auto regressions “I love machine learning in ___”

AE: auto encoder “I love [mask] in NLP”

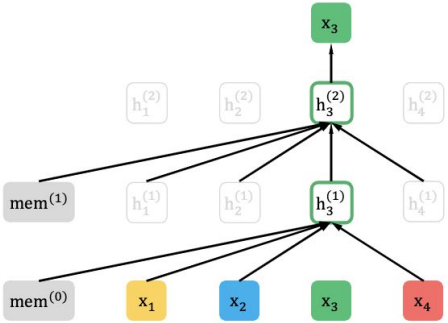
PLM: “machine in love I NLP __”



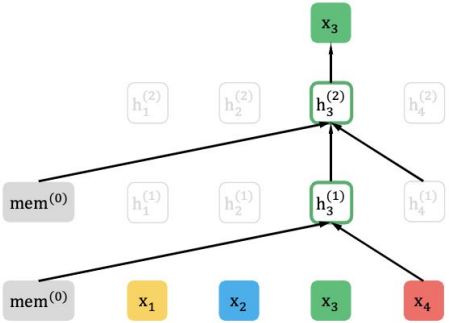
Factorization order: 3 → 2 → 4 → 1



Factorization order: 2 → 4 → 3 → 1



Factorization order: 1 → 4 → 2 → 3



Factorization order: 4 → 3 → 1 → 2

XLNet : How attention mask works in detail

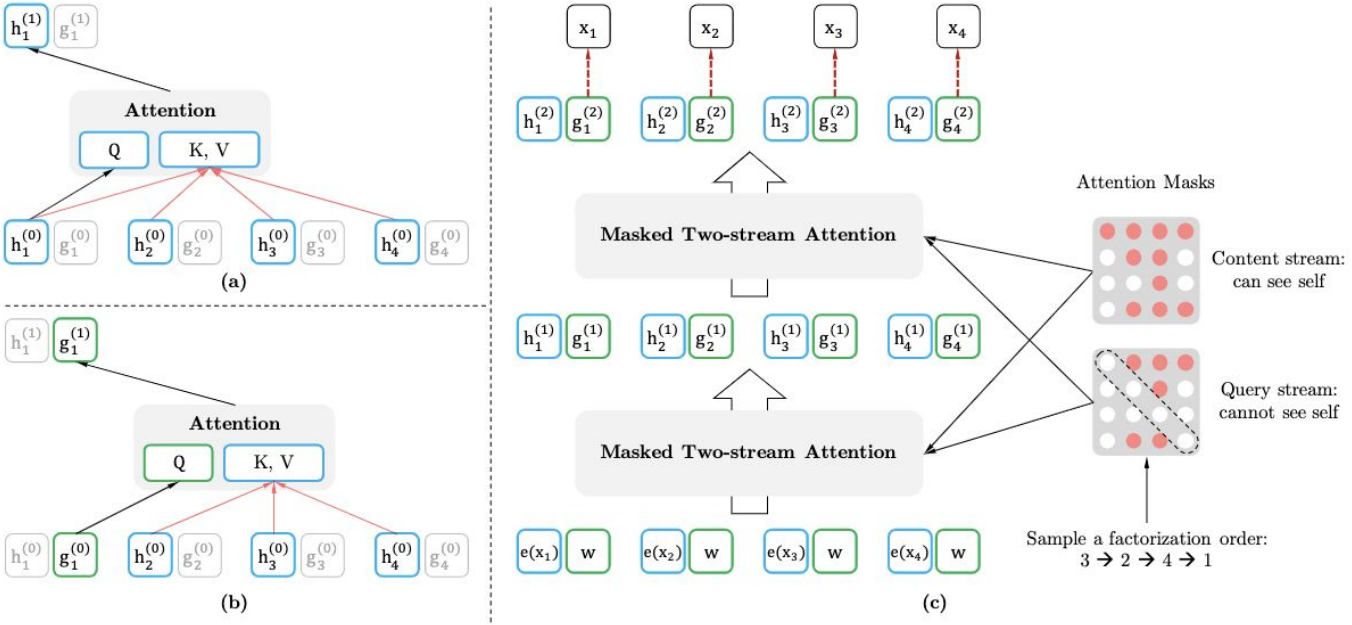


Figure 1: (a): Content stream attention, which is the same as the standard self-attention. (b): Query stream attention, which does not have access information about the content x_{z_t}. (c): Overview of the permutation language modeling training with two-stream attention.

Albert: an improved model from BERT

1. Factorized embedding parameters

Bert embedding parameters:

hidden layer size * embedding size * embedding matrix = billions

Albert: lower dimension in hidden layer matrix

Albert: an improved model from BERT

2. Cross layer parameter sharing

Feed forward network and attention parameters are all shared

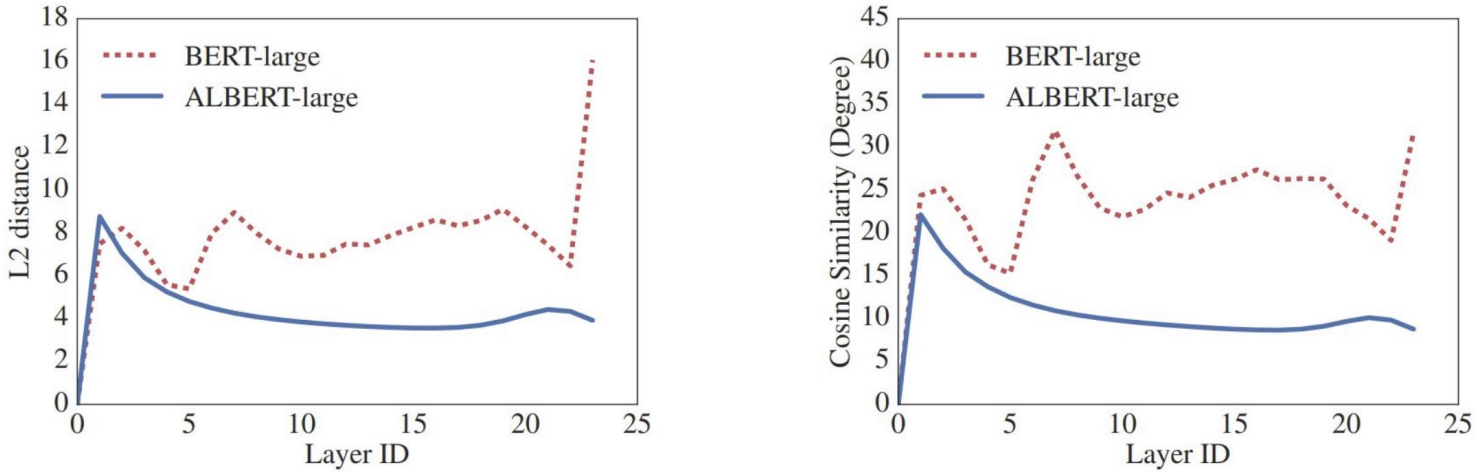


Figure 2: The L2 distances and cosine similarity (in terms of degree) of the input and output embedding of each layer for BERT-large and ALBERT-large.

Albert: an improved model from BERT

3. Use sentence order prediction instead of Next sentence prediction

- Improved the efficiency of the model
- Proved by showing better resulting data

Result of our experiment

Different max sequence length

max_lenght	F1	Has ans F1	No ans F1	Time per epoch
64	\	\	\	\
100	71.64	55.98	87.28	0:39
128	**77.26**	**69.57**	**85.94**	**0:52**
256	78.98	70.88	82.51	1:39
348	81.67	74.08	80.27	1:57

Different epoch

max_lenght	Epoch	F1	Has ans F1	No ans F1
128	2	74.33	67.27	85.34
128	5	75.26	70.2	86.3
128	10	75.99	70.89	87.33

Different Albert pretrained

model	max_lenght	Run time	F1	Has ans F1	No ans F1
Base	128	0:52	77.26	69.57	85.94
large	128	2:36	79.03	71.28	86.75
xlarge	128	9:05	80.23	72.33	88.04

Challenges

- Computation time
- GPU memory size (Cuda out of memory).
- RuntimeError: write(): fd 28 failed with No space left on device.
  - Increase save_steps.
  - Use smaller batch size.
- Disconnect from colab: Need to restart the training, and feature conversion
- Buffered data was truncated after reaching the output size limit.
  - Limiting output logging or any other similar mechanism won't work
  - Save_steps and continue train from checkpoint