“Attention” Class! Let’s solve some reading comprehensions

dewang, murri, pnorouzi, oshay

Dataset

RACE: Large-Scale ReAding Comprehension Dataset From Examinations

Contains:

Train Dataset : 25139 Articles, each with ~ 4 Questions

Dev Dataset: 1391 Articles

Test Dataset: 1409 Articles

Why is the dataset challenging and interesting?

- Race is created by domain experts to test student's reading comprehension skills, consequently requiring non trivial challenging techniques
- A significant gap between state of the art and human performance.
- RACE has a variety of question types. Here are some samples:
  - What is the best title of the passage(summarization)
  - What was the author's attitude towards the industry awards?(Inference)
  - Which one of the following is WRONG according to passage? (Deduction)
  - The first postage stamp was made by? _ (Context Matching)

Model	Report Time	Institute	RACE	RACE-M	RACE-H
Human Ceiling Performance	Apr. 2017	CMU	*94.5*	*95.4*	*94.2*
Amazon Mechanical Turker	Apr. 2017	CMU	73.3	85.1	69.4
Dual Co-Matching Network (DCMN) (ensemble)	Mar. 2019	SJTU & CloudWalk	**74.1**	**79.5**	**71.8**
Option Comparison Network (OCN) (ensemble)	Mar. 2019	Pattern Recognition Center, WeChat AI, Tencent Inc	73.5	78.4	71.5
Dual Co-Matching Network (DCMN)	Mar. 2019	SJTU & CloudWalk	72.3	77.6	70.1
BERT_LARGE	Feb. 2019	Tencent AI Lab	72.0	76.6	70.1
Option Comparison Network (OCN)	Mar. 2019	Pattern Recognition Center, WeChat AI, Tencent Inc	71.7	76.7	69.6
BERT_LARGE	Jan. 2019	River Valley High School, Singapore	67.9	75.6	64.7
Reading Strategies Model (ensemble)	Oct. 2018	Tencent AI Lab & Cornell	66.7	72.0	64.5
BERT_BASE	Jan. 2019	River Valley High School, Singapore	65.0	71.7	62.3
Reading Strategies Model	Oct. 2018	Tencent AI Lab & Cornell	63.8	69.2	61.5
GPT	June 2018	OpenAI	59.0	62.9	57.4
Convolutional Spatial Attention (ensemble)	Nov. 2018	Joint Laboratory of HIT and iFLYTEK Research	55.0	56.8	54.8
BiAttention (MRU) (ensemble)	Mar. 2018	Nanyang Technological University & Institute for Infocomm Research	53.3	60.2	50.3
Dynamic Fusion Networks (ensemble)	Nov. 2017	MSR & CMU	51.2	55.6	49.4
Convolutional Spatial Attention	Nov. 2018	Joint Laboratory of HIT and iFLYTEK Research	50.9	52.2	50.3
BiAttention (MRU)	Mar. 2018	Nanyang Technological University & Institute for Infocomm Research	50.4	57.7	47.4
Hierarchical Co-Matching	June 2018	Singapore Management University & IBM Research	50.4	55.8	48.2
Dynamic Fusion Networks	Nov. 2017	MSR & CMU	47.4	51.5	45.7

Approaches

- Non Deep Learning: A Logistic Regression with average GLoVE embeddings
- Baseline Deep Learning: A CNN with GloVE embeddings
- BERT Base
- BERT Large

Data Hypothesis - 1

Here are some things which we believe that a good model should be able to do:

- **Model should be able to pick the synonym option given a passage with an explicit answer.**

So his teachers didn't like him, and nor did his classmates play with him.Peter often slept in class because his heart was not in school. He almost gave himself up.One day....

Peter always failed in exams because he?

- Went gambling
- Talked in class.
- Didn’t answer
- **Gave himself up**

If we change the option from **gave himself up** to a synonym like **“stopped trying”**, the model should still be able to predict it.

Data Hypothesis - 1

Here are some things which we believe that a good model should be able to do:

- **Model should be able to pick the synonym option given a passage with an explicit answer.**
  - BERT got it correct because the attention mechanism related stopped trying to gave himself up.
  - Logistic Regression predicted **gave himself up** correctly but failed to find the correct answer when we substituted a synonym
    - It instead chose **talked in class** as the correct answer.
  - CNN correctly predicted the right answer in each case.
    - This means that it can understand synonyms in some settings

Data Hypothesis - 2

Here are some things which we believe that a good model should be able to do:

- **Model should identify a subject action relationship.**

Alice likes sunny weather just like today. She wants to know what the weather will be like tomorrow. She's going to have a picnic.This is what the reporter is saying....

What is she going to do ?

- She's going to work
- **She's going to have a picnic**
- She's going to watch TV
- She's going to wash clothes

If we change **She's going to have a picnic** to **She's going to wash clothes** in the passage, the model should change it's answer.

Data Hypothesis - 2

Here are some things which we believe that a good model should be able to do:

- **Model should identify a subject action relationship**
  - BERT Works because Attention relates the subject to the action.
  - Logistic Regression failed to predict **She's going to wash clothes** when we changed the corresponding action in the article
    - It instead still chose **She's going to have a picnic** as the correct answer.
  - CNN did not even get the correct answer in the first case!
    - In both cases it predicted **She's going to work**!

Data Hypothesis - 3

Here are some things which we believe that a good model should be able to do:

- **Model should understand chronology of actions**

  Alice went to the store after school. Then Alice went home...

  Where is Alice now?

  - Store
  - School
  - **Home**
  - Picnic

  The model should be able to capture chronological options and predict **Home**.

Data Hypothesis - 3

Here are some things which we believe that a good model should be able to do:

- **Model should understand chronology of actions.**
  - BERT sometimes works and sometimes it does not, maybe attention is not all you need.
  - Logistic Regression fails to predict chronological order of actions.
    - In the question in the previous slide it predicted **picnic** as the correct answer.
  - CNN wrongly predicts **School**
    - This means that CNN model did not understand chronological order of actions.

Data Hypothesis - 4

Here are some things which we believe that a good model should be able to do:

- **Correct identification of Proper Nouns**

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University...

What city did Tesla move to in 1880?

- Gospic
- **Prague**
- Czech
- Ferdinand

If we change **Prague** to **Delhi**, the model should change it's prediction to **Delhi**.

Data Hypothesis - 4

Here are some things which we believe that a good model should be able to do:

- **Correct identification of Proper Nouns**
  - BERT works here because attention mechanism relates the place no matter what is placed there. In case we put a city not present in options, it predicts Gospic, which might be second in list of attention.
  - Logistic Regression fails to identify proper nouns in most cases.
    - In the question in the previous slide it predicted **Ferdinand** as the correct answer doesn't matter if we changed the location or not.
  - CNN correctly predicts **Delhi** when we change the location!
    - It correctly identifies proper nouns.

Data Hypothesis - 5

Here are some things which we believe that a good model should be able to do:

- **Model should not answer based on Question/Passage sentence similarity:**

In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived.....

What city did Tesla move to in 1880?

- Gospic
- **Prague**
- Czech
- Ferdinand

If we add a sentence **Tomato moved to Czech in 1880**, this sentence is close to the question in euclidean space, but the model should not change it's prediction based on that.

Data Hypothesis - 5

Here are some things which we believe that a good model should be able to do:

- **Model should not answer based on Question/Passage sentence similarity:**
  - BERT fails here, Well... Attention is not all you need I guess.
  - Logistic Regression fails to succeed in distinguish between similar sentences.
    - In the question in the previous slide it predicted **Ferdinand** as the correct answer doesn't matter if we added a similar sentence or not.
  - CNN surprisingly performs pretty well! It correctly doesn't change its answer!
    - It means that it doesn't just choose the right answer solely because of similarity!

Data Hypothesis - 6

Here are some things which we believe that a good model should be able to do:

- **Model should not make up answers**

Charlie likes dogs. Charlie likes to read and sing.

What's Charlie's favourite color?

- Red
- Blue
- Yellow
- **Not enough information**

The model **should not try to predict a color** here.

Data Hypothesis - 6

Here are some things which we believe that a good model should be able to do:

- **Model should not make up answers**
  - BERT FAILS because it tries to predict colors. People have tried getting around that by adding a column call is_answerable and train the model again using that information.
  - Logistic Regression fails in this task as well
    - It chose **Yellow** for some reason!!
  - CNN not surprisingly also fails in this test.
    - It chose **Blue** for some reason!

Performance on Data Hypothesizes

	Understand Synonyms	Subject Action Relationship	Chronology of Actions	Identify of Proper Nouns	Should not Answer from Similarity	Should not make up answer
Bert Base						
CNN						
Logistic Regression						

Analysing some more BERT failures

- In our setting because of memory constraints we fixed the max sequence length to 200, So for long passages BERT tends to fail.

- To get around this we tried to summarize the passage using gensim summarizer module and pass the passage again.

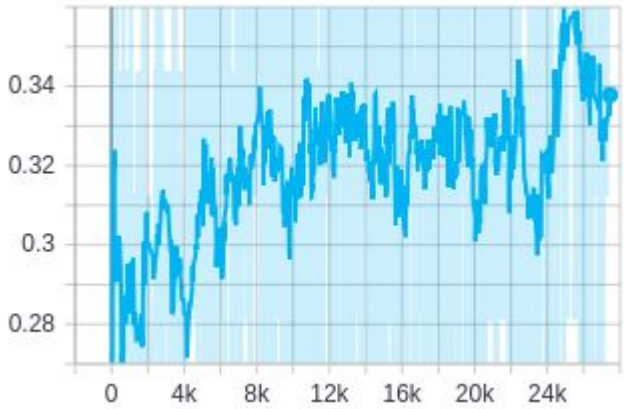
- It certainly makes predictions for certain long passages better but at the same time fails some other long passages because the information required to learn those is lost.

One Solution for Long Passages

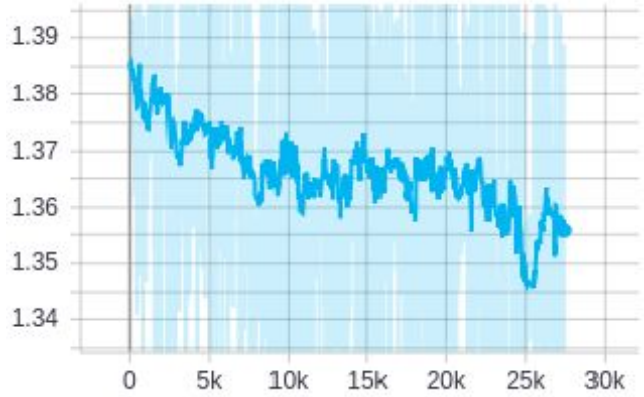
- Our Bert model had a max sequence of 200 but some the passages are more than 1000 words!
  - For one of the long passages we got none of the questions correctly due to the small max sequence length.

- To solve this issue, we used **gensim text summarizer** on the passages that were longer than a set threshold.
  - For the passage explained above, this method helped us to get all of the questions related to that passage correctly
  - This is not a total fix for this problem because by summarizing the text we may be losing inherently important information about the passage that may help with logical inference.

Results - Logistic Regression Plots

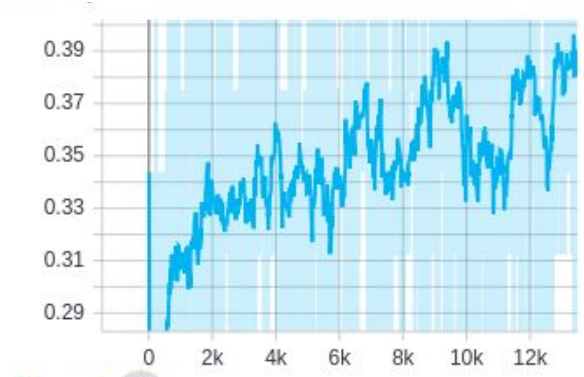


Train Accuracy

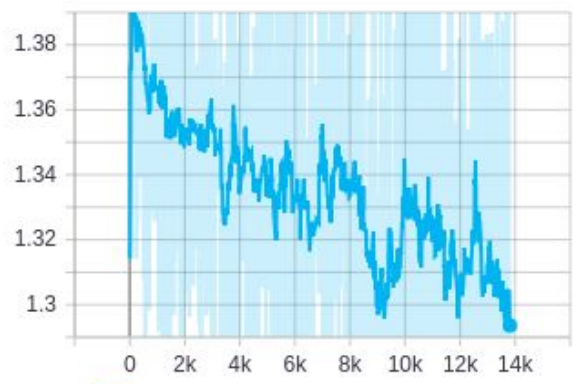


Loss

Results - CNN Plots

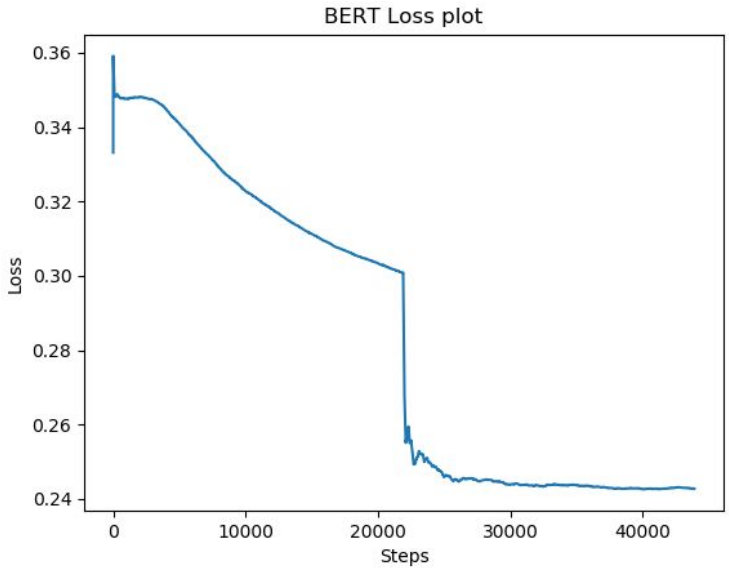


Train Accuracy



Training Loss

Results - Bert Base Plot



Training Loss

Performance on the Dataset

	Final Train Accuracy (%)	Final Dev Accuracy (%)	Final Test Accuracy (%)
Bert Base	59.80	57.35	56.70
CNN	39.33	35.52	34.09
Logistic Regression	32.41	29.10	28.22