Long Range Dependencies in Biomedicine

Group Fantastic Five

Koushik Sai Achyuth Ayila
Nagacharan Vemula
Hari Sai Charan Challa
Ashish Wale
Maaz Khazi

Mentors

Mihir Parmar

Divij Handa

MOTIVATION

- Surge in machine-readable text within the medical field particularly in Electronic Health Records (EHRs).
- Growing demand in the real-world healthcare scenarios for LLMs to handle excessively longer and more complex documents.
- Several LLMs like BioGPT, BioMedLM, GatorTRONGPT and MedPaLM have been developed in these domains although their efficacy is confined to smaller texts.

PROBLEM STATEMENT

- The limited input token length window of LLMs becomes a bottleneck when dealing with lengthy contexts like EHRs (Medical data). This project is dedicated to addressing the challenge of capturing long contexts and long-range dependencies in biomedical data, considering the constraint imposed by the fixed token length of LLMs thus improving their performance.

APPROACH - SELECTIVE CONTEXT

- Employed a method called **Selective Context** that uses self-information to selectively filter out less informative content.

- Uses a base language model to figure out self information of lexical units. Lexical units can be a token or a phrase or a sentence.

- Helps make the context simpler and more efficient for large language models to understand and work with.

APPROACH - SELECTIVE CONTEXT

- **Self Information** represents a concept within information theory that measures the quantity of information conveyed by an event. In the context of language modeling, the event corresponds to a single step of generation, specifically a token.

- Given a context {x0, x1, ..., xn}, Self-information of the a token x_i, can be calculated as $\boldsymbol{I(x_i) = -\log_2 P(x_i \mid x_0, x_1, \ldots, x_{i-1})}$, where $P(x_i \mid x_0, x_1, \ldots, x_{i-1})$ is the output probability of token x_i extracted from the chosen base model(Ex:GPT-2).

- Tokens with higher probabilities have lower self-information. We want to remove lower perplexity tokens or tokens with higher output probabilities as these tokens have less informative context(Commonly observed tokens).

APPROACH - TOKEN SELECTION

- The selective context method employs a percentile-based filtering strategy to dynamically choose the most informative content.

- Initially, the tokens are ranked in descending order based on their self-information values. Following that, it determines the p-th percentile of self-information values among all tokens based on the chosen p.

- Then, it selectively keeps tokens whose self-information values are greater than or equal to the p-th percentile, forming a filtered context C' from C.

APPROACH - CHOICE OF P/REDUCTION_RATIO

- In our approach, we have decided to set p dynamically for each input such that the token length of compressed context is less than the maximum limitation of the encoder-decoder model(1024).

- We are setting,
  p = (1 - (max_token_length_of_model/instance_token_length))*100
  = (1 - (1024/instance_token_length))*100.

APPROACH - SELECTIVE CONTEXT API

- Used the Selective Context API to compress the input instances in our data, enabling two key operations:
    - Computing the self-information values.
    - Filtering out less informative context using the self information values.

- Iteratively converted all the input data instances into compressed instances.

- Used different reduction ratios(‘rr’) for different data instances based on their token length.

APPROACH - USAGE OF THE API

```
sc = SelectiveContext(model_type='gpt2', lang='en')

def filter_content(initial_df):
        df_modified = initial_df.copy()
        for i in range(len(initial_df)):
            sc_encoding = sc_tokenizer(initial_df["input"][i],
add_special_tokens=False, return_tensors='pt')
            sc_input_ids = sc_encoding['input_ids'].squeeze().tolist()
            if len(sc_input_ids) > 1024:
                rr = 1 - (1024/len(sc_input_ids))
                context, reduced_content = sc(initial_df["input"][i],
reduce_ratio=rr, reduce_level="token")

                df_modified["input"][i] = context

        return df_modified
```

RESULTS - BASELINE APPROACH

- Baseline approach: To truncate the input data to match each model's maximum input token length and then assess how each model performs.
- Models Used: cogint/in-boxbart , razent/SciFive-large-Pubmed_PMC

Dataset	In-BoXBART(Accuracy)	Scifive(Accuracy)
Smoking Challenge Data 2006	56.7308	60.5769
Obesity Challenge Data 2008	71.86	71.3496
Heart Disease 2014	48.3307	51.1526
Cohort Selection 2018	57.7818	57.424
Assertions Challenge Data 2010	68.2012	68.0832
Temporal Relations 2012	54.1315	56.0394
ADE 2018	18.1421	19.5892

RESULTS - SELECTIVE CONTEXT APPROACH

- Following are the results obtained upon training the models(cogint/in-boxbart , razent/SciFive-large-Pubmed_PMC) with the compressed context,

Dataset	In-BoXBART(Accuracy)
Smoking Challenge Data 2006	60.5769 ↑
Heart Disease 2014	50.1987 ↑

Dataset	Scifive(Accuracy)
Smoking Challenge Data 2006	60.5769

ANALYSIS

- As the accuracy is increased using the selective context approach for InBoXBART on Smoking Data and Heart Disease Data, we took an instance of Smoking data and analysed the compressed context of that instance.
- Full example link : [Selective Context Example]

305757070 ELMVH 74973293 2973118 ~~11/27/2003~~ 12:00:00 AM dehydration DIS Admission Date : ~~11/27/2003 Report Status : Discharge Date : 11/28/2003 ****** DISCHARGE ORDERS *****~~* MEEDBELB , LIND E 869-08-73-5 B09 Room : 23Q-929 Service : CAR DISCHARGE PATIENT ON : ~~11/28/03 AT 02:00 PM CONTINGENT UPON Not Applicable~~ WILL D / C ORDER BE USED AS ~~THE D / C SUMMARY :~~ YES Attending : KOTESKISMANHOUT , AYAN , M.D. CODE STATUS : Full code DISPOSITION : Home ~~DISCHARGE MEDICATIONS :~~ COLACE (DOCUSATE SODIUM) 100 MG PO BID FENTANYL (PATCH) 25 MCG / HR TP Q72H Alert overridden : Override added ~~on 11/27/03~~ by CRAMPKOTE ~~, LINE , M.D. POSSIBLE ALLERGY (OR SENSITIVITY) to~~ NARCOTICS , PHENYLPIPERIDINE POTENTIALLY ~~SERIOUS~~ INTERACTION : CITALOPRAM HYDROBROMIDE and ~~FENTANYL~~ CITRATE Reason for override : takes at home w / o problem LASIX (FUROSEMIDE) 40 ~~MG PO~~ monday ~~, wednesday , friday~~ ZESTRIL ~~(LISINOPRIL) 20 MG PO QD~~ HOLD IF : sbp < 100 Override Notice ~~: Override added on 11/27/03 by CRAMPKOTE , LINE , M.D.~~ on order for KCL IMMEDIATE RELEASE PO (ref # 16679329) ~~POTENTIALLY SERIOUS INTERACTION : LISINOPRIL and POTASSIUM CHLORIDE Reason for override :~~ aware TOPROL XL (METOPROLOL (SUST. REL.)) 25 MG PO QD HOLD IF : hr < 55 Food / Drug Interaction

ANALYSIS

- From the analysis, we observed that this mechanism is able to filter out the unimportant information such as Dates, Stop words etc.. and it is able to retain the relevant information related to Drugs that is helpful in predicting the class.
- Using Selective context with the Scifive did not improve the accuracy. We found that the scifive model is predicting the same label “UNKNOWN” for all the inputs. This is because there is an extreme class imbalance, with a significant 63.40% class labels falling under “UNKNOWN”. Dealing with this class imbalance can improve the accuracy of the model.

REFERENCES

- Li, Yucheng. "Unlocking Context Constraints of LLMs: Enhancing Context Efficiency of LLMs with Self-Information-Based Content Filtering." *arXiv preprint arXiv:2304.12102* (2023).
- https://huggingface.co/cogint/in-boxbart
- https://huggingface.co/razent/SciFive-large-Pubmed_PMC