Track Details

Company Name : CACTUS COMMUNICATION

# of tracks you are sponsoring : 1

Prize Money amount details you will be sponsoring for each track:

- Winner : INR 15000
- First runner up : INR 10000
- Second runner up (optional) : INR 5000
- Consolation prize (optional) : NA

Note – We are also looking to hire interns/future engineers and linguists/scientific researchers and the hackathon is a good platform for folks who want to showcase their work directly to us. They can be offered direct interview for the job roles post hackathon.

Track Name : Making Scientific Research Accessible Using AI and Big Data

Problem Statement #1: Netflix for Researchers

Overview

One of the key problems a researcher faces is to find the right paper to read while doing literature survey or in general to keep up to date in their field/area of research. While in some areas the papers are flooded (based on interest peaks over time), sometimes its difficult to find the right paper and only after going through the paper which requires significant time and effort, one realizes that this paper is not useful to them.

Sounds similar to what a certain movie aggregator platform tried to solve but for movies? You are on the right track then (pun intended!)

The idea is to build a recommender system for papers.

We have already built something of this sort in our ecosystem of products and for reference, you can check it out on web and even download the app on playstore/Appstore to understand the end use experience a bit.

https://discovery.researcher.life/

Expected outcomes

While there are many ways to build a recommender system, what our solution does is focuses on what we call as concepts. Our proprietary system generates concept tags for each paper and then we use them to recommend papers in a structured way (apart from many other params which for obvious reasons we can't reveal :D)

This usually falls in a very traditional content recommendation logic. While you can figure out a way to do this, what we would also want for you to think and build a prototype which can recommend papers using collaborative/social media networks/hybrid approaches and if you can one up on our system (so that we can directly hire you and send you this meme instead from a famous Indian movie)

Technical outcomes –

1. A working mvp (app/site) or a demo UI which can demonstrate the search and recommend feature
2. Recommending papers for sample personas/users and walking through the algorithm. For example,
    a. Let's say you are the father of deeplearning but have been busy with making companies succeed that you are out of sync of what is happening in the field of semi-supervised learning. Which would be the top 5 papers to suggest him.
    b. You are a great pyshciatrist specializing in the field of hypnotherapy. You want to know if learning other researchers in the field have come up with something new and want to see if any cross-field papers are published.

3. [Optional]Any additional layers/topics which you can pick up which goes into making reading paper experience better. For example – detecting an emerging concept before it becomes big/mainstream – COVID-19 being the best example.

Dataset/ References
https://openalex.org/
https://paperswithcode.com/paper/a-scalable-hybrid-research-paper-recommender
https://github.com/akanakia/microsoft-academic-paper-recommender-user-study
https://paperswithcode.com/paper/hierarchical-bi-directional-self-attention
https://scholarbank.nus.edu.sg/handle/10635/146027
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8378599/
https://pubmed.ncbi.nlm.nih.gov/
https://arxiv.org/abs/2008.13538
https://www.db.soc.i.kyoto-u.ac.jp/~sugiyama/SchPaperRecData.html
https://arxiv.org/abs/1706.03428

Problem Statement #2: Raiders of the lost Manuscript

Overview

There are many ways in which a researcher can present their work. A poster, a case report, some conference proceedings, just some abstract about an research idea, their dissertations, some clinical trials and observations. The list goes on. As a scientific editing business unit, we get anything and everything to be worked on right from that 100 word abstract to 10000 words heavy dissertation.

As there are countless of online platforms where people publish their research and in different format, it becomes difficult to identify what type of manuscript it is without someone going through the document manually and deciding/tagging it as one of many. Sounds familiar to a certain group of unique document(s) used for identification of an individual, doesn't it?

Well the idea is to not only detect what type of manuscript document it is but to also identify the structure it contains.

For example, a typical IEEE computer science paper follows a 2 column structure with a fixed sequence of Title, authors, abstract, keywords, introduction and so on.

What if we can also identify these structures and extract information/text per section? Well that is the next logical step given we are dealing with different types of media – PDFs, Infographics, Figures, Aniamted GIFs.

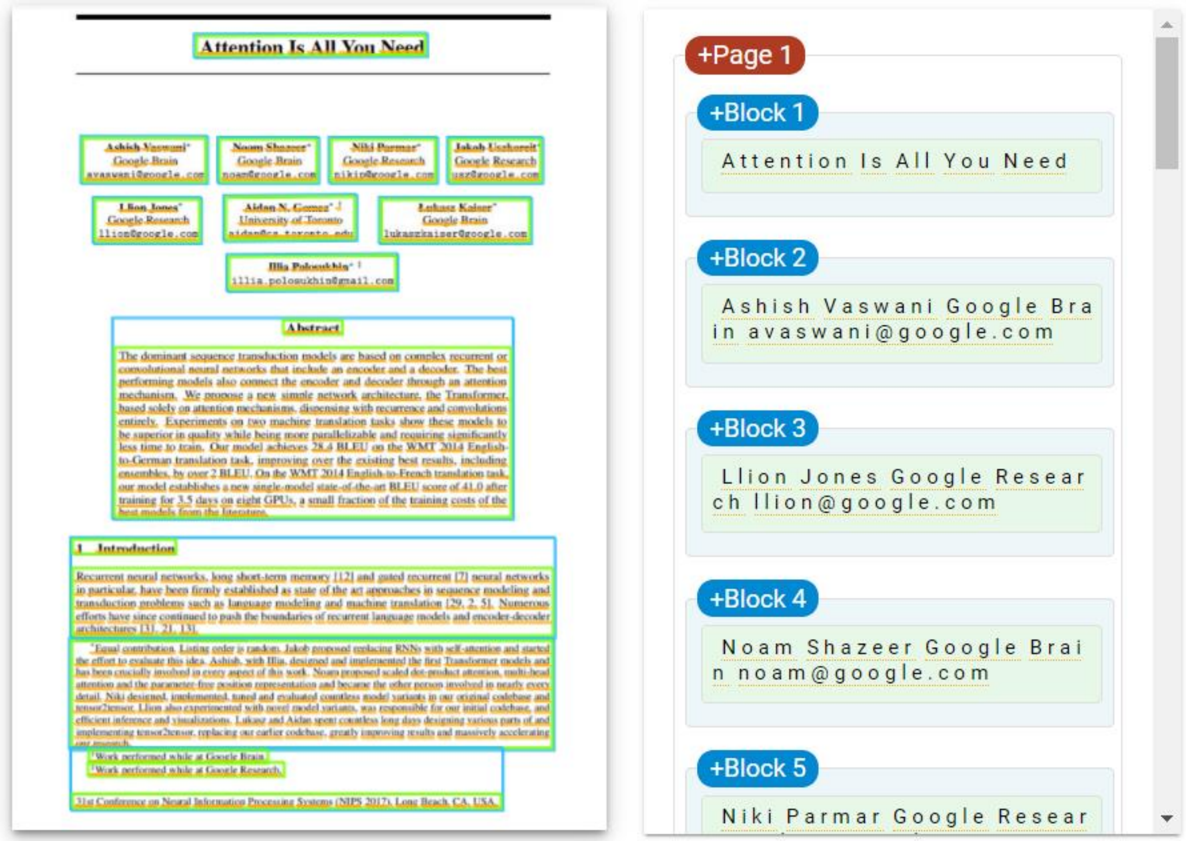
The idea over here is then simple –

We need to find how many manuscript doc types can we detect and structure. A few examples which come to our mind are :

- Abstracts
- Case reports
- Clinical trials
- Conference proceedings
- Thesis/Dissertations

You can also start with one type of doc, try to figure out a way to identify the structure of the document and then also extract text in proper format from it.

A rough example of this would be with this nifty vision api by google https://cloud.google.com/vision (which can get expensive and clunky in no time so using it just as a way to explain what we are looking for)



Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

1 Introduction

Recurrent neural networks, long short-term memory [12] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [29, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [31, 21, 13].

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

+Page 1

+Block 1

Attention Is All You Need

+Block 2

Ashish Vaswani Google Brain avaswani@google.com

+Block 3

Llion Jones Google Research llion@google.com

+Block 4

Noam Shazeer Google Brain noam@google.com

+Block 5

Niki Parmar Google Resear

f2.JPG

Now this can be a blend of both a computer vision and NLP problem. You can choose many ways to reach a result even by just following a singular approach. This is where it will be interesting to see if you can also try multiple approaches, think like a researcher and benchmark for a same type of document, which approach is working the best.

Expected technical outcomes

- A simple API or UI which can take document (pdf/doc) as an input and give information about the type of manuscript and extract information from it and give it in a structured manner.
- Some experimental approaches if documented can also be showcased with outcomes per methodology.
- Covering as many manuscript types as possible if one type you feel is solved (scalable)
- Create your own toy dataset and submit that as well

Dataset/References

https://ieeexplore.ieee.org/document/8270123
https://link.springer.com/book/10.1007/978-3-030-86549-8
https://www.editage.com/insights/how-can-i-identify-the-article-type-for-a-manuscript-i-am-reviewing
https://authorservices.taylorandfrancis.com/publishing-your-research/writing-your-paper/different-types-of-research-articles/
https://www.researchgate.net/publication/226300537_Document_Structure_and_Layout_Analysis
https://arxiv.org/pdf/1910.03678.pdf
https://paperswithcode.com/task/document-layout-analysis

Problem Statement #3: Open Problem Statement

Overview

While we understand that there may be something which doesn't fit your expertise, this section we have kept as an open idea box where if you feel there is an interesting problem in the researcher ecosystem which you have faced and know can be solved using AI and Big Data, feel free to form your own problem statement and work on it. Make sure you are also sourcing your datasets from proper places and you will need to share that data/data source with us.

Some open ideas –

1. Plagiarism Detection (which works also for images)
2. Multipaper summary – Can you condense 5 papers of same research area into one?
3. Emerging Concept Detector – Can you predict which topic of research will become important in next month/year/decade? Finding interest in emerging topic and a dashboard to track the trends would surely be interesting.

And the list goes on!!

Expected outcomes

While we do encourage open innovation, if you take up this problem statement, make sure you are formulating your idea and use case well and get it validated from our mentors who will be around on Day 1 itself for getting maximum time to build. Once you get the go ahead from them, you can start to build on it.

Note: It is important if you choose to work on your own problem statement to get them validated first with any of CACTUS staff present as mentors. Failure to do so will dent your final evaluation score.

Submission Guidelines

- Code hosting can be in a private gitrepo which needs to be shared with our mentors.
- Presentation/Documentation should also be uploaded in the same repo so that we can download and refer it later.
- We will let you know over if we feel the repo should be made public/kept private once the hackathon is over.

Mentor(s):

For everything and in general (even for generic track)

1. Parth Agrawal (AI Evangelist/Product Manager) - https://www.linkedin.com/in/htrap94/
2. Mayank Raj (Associate Director Of Engineering)- https://www.linkedin.com/in/mayank9856/

Will be present in track discord group to help participants out

1) Sartaj Pannu (Senior Software Engineer – BigData) - https://www.linkedin.com/in/sartaj-pannu-47935780/
2) Pooja Mehta (Senior Software Engineer – BigData) - https://www.linkedin.com/in/pooja-mehta-9986b913a/
3) Krishna Mishra (Senior Software Engineer – NLP) - https://www.linkedin.com/in/krishnakumar-mishra-a03278b9/
4) Kunal Sawhney (Engineering Team Lead – Big Data and NLP) - https://www.linkedin.com/in/kunal-sawhney-607aa3bb/
5) Parth Chudasma (Software Engineer – NLP) - https://www.linkedin.com/in/parthchudasama12/

Evaluation :

For evaluation, the mentor pool mentioned above will be there. I’ll confirm if our CTO, Mr. Nishchay Shah (Linkedin Profile) will be joining or not.

Criteria

Weightage to be given to

a. Algorithmic Approach (20%)
b. Code (20%)
c. Presentation(20%)
d. Demo(20%)
e. Scalability(20%)

Signature: Parth Agrawal (On behalf of CACTUS COMMUNICATION)