MuLD: The Multitask Long Document Benchmark

G Thomas Hudson, Noura Al Moubayed

Scan me for the Github repo

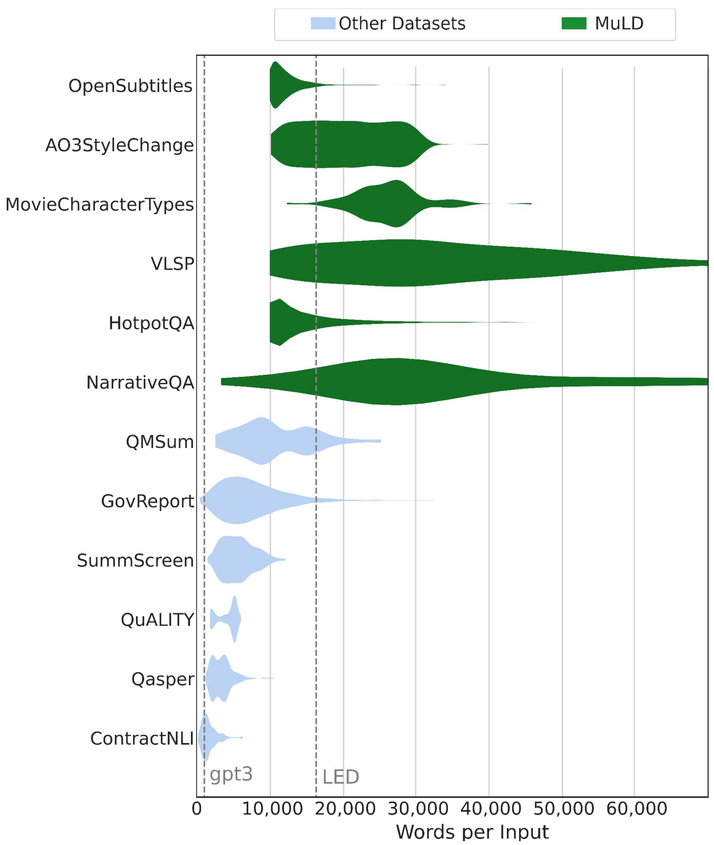
Durham University

Problem

The field of Natural Language Processing (NLP), has seen breakthroughs in the use of increasingly large pretrained language models, driven by competition on influential benchmarks such as GLUE [1].

Recently, efficient transformer models such as Longformer [2], and Reformer [3] have explored techniques to allow transformers to operate on much longer sequences. However most of the datasets used to evaluate these models only use a few thousand tokens.

Our benchmark: MuLD, consists of 6 tasks where the inputs are at least 10,000 tokens.



Comparison of dataset lengths of MuLD tasks and other datasets

Tasks

MuLD consists of 6 tasks chosen to span a variety of dataset sizes, genres, and formulations, and are created by filtering, extending, or modifying existing NLP datasets.

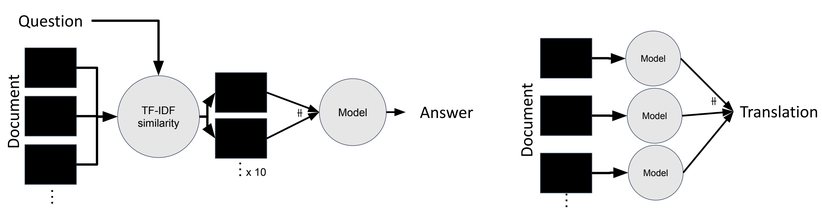
The benchmark covers a wide variety of task types including translation, summarization, question answering, and classification. Additionally there is a range of output lengths from a single word classification label all the way up to an output longer than the input text.

Task	Description	Example
NarrativeQA	Dataset of user-submitted questions regarding the plot of a movie or novel. We filter out any documents shorter than 10,000 tokens.	**Input:** How is Oscar related to Dana? A high AERIAL SHOT of the island features the Statue of Liberty prominently in the foreground[...] DANA Frank, do you think you could give me a hand with these bags? FRANK I'm not a doorman, Miss Barrett. I'm a building superintendent[...] **Output:** her son
HotpotQA	Question answering dataset requiring information from multiple Wikipedia pages to answer - a test of multihop reasoning. The original dataset used snippets of Wikipedia articles, which we expand into the full text.	**Input:** Were Scott Derrickson and Ed Wood of the same nationality? Doctor Strange is a 2016 American superhero film[...] Scott Derrickson (born July 16, 1966) is an American filmmaker. He is best known for directing the films The Exorcism of Emily Rose, Sinister, Deliver Us from Evil, and Doctor Strange[...] Edward Davis Wood Jr. (October 10, 1924 - December 10, 1978) was an American filmmaker, actor, and author[...] **Output:** yes
Character Archetype Classification	Classification dataset where the task is to identify the archetype of a character (Hero, Villain, etc.) based on reading the movie script. We gather this dataset based on the methodology of [4].	**Input:** Indiana Jones[...] INDY No. Barranca looks evilly at Indy's hand upon him. Indy releases him and smiles in a friendly way. INDY We don't need them.[...] **Output:** Hero
Open Subtitles	A translation dataset based on aligned subtitles of movies and tv shows from the crowdsourced site opensubtitles.org.	**Input:** 957 was a big year. The Russians put that Sputnik into outer space. The Dodgers played their last game at Ebbets Field and said goodbye to Brooklyn.[...] **Output:** 957 war ein bedeutendes Jahr. Die Russen schossen ihren Sputnik ins All. Die Dodgers spielten zum letzten Mal in Ebbets Field und sagten Brooklyn Adieu.[...]
AO3 Style Change Detection	Dataset where the task is to identify where the author changes in a document constructed from the work of multiple authors. Introduced as a PAN21 shared task [5], we use fanfiction stories from the website archiveofourown.org instead of the shorter StackExchange comments.	**Input:** John's stomach had a new home again. It now settled into his throat, painfully dry for want of the man before him.[...] Right, but I still reckon I'll keep pretending you don't." "Don't mention it," Sherlock responded with a soft passion he couldn't keep out of his voice. The words came out in a murmur. -" he continued, delighted when John joined in to harmonize flawlessly in the latter half of phrase[...] **Output:** 1,[...],1,2,2[...]
VLSP	A version of the Scientific Papers summarization dataset of (abstract, paper) pairs [6], where we use only long theses rather than excluding them.	**Input:** ## Chapter 1 Basic Definitions and Concepts ### 1.1 Basics of simplicial complexes Let @xmath , and @xmath denote the subsets of @xmath of size @xmath . A collection @xmath of subsets of @xmath is called a (finite abstract) simplicial complex if it is closed under inclusion, i.e. @xmath implies[...] **Output:** This thesis focuses on algebraic shifting and and its applications to f-vector theory of [...]

Baselines

We experiment with two models: T5 representing 'standard' transformer models, and Longformer representing 'efficient' transformer models. As our documents are too long to be processed by either of these models on reasonable hardware, we devise chunking techniques for each task:

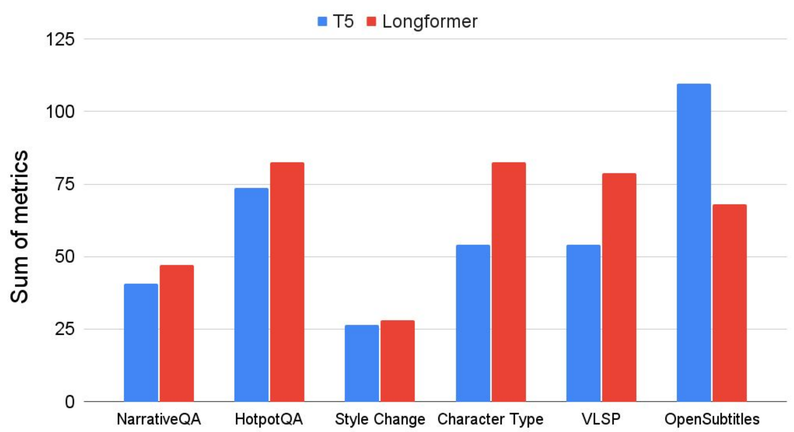
QA Chunk the document, then use TF-IDF between chunks and the question to pick the most similar 10 chunks to feed into the model.
Translation Pass each chunk into the model then concatenate the result.
Style Change Detection Train a classifier to check if paragraph pairs are written by the same author. This is applied repeatedly to identify the author of each paragraph.
Character Archetype Classification Select chunks containing the first, last, and most frequent mention of the character to pass into the model.
Summarization Summarize the text from the first introduction section onwards.



i) Question Answering *ii) Translation*

Illustration of the chunking methods for two of the MuLD task types.

Results



The Longformer model outperforms the T5 model across many of the tasks, suggesting models which are able to make use of a longer context perform well on our benchmark.

[1] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo, 'Segformer: Simple and efficient design for semantic segmentation with transformers'
[2] Beltagy, I., Peters, M. E., and Cohan, A. (2020). 'Longformer: The long-document transformer'. arXiv preprint arXiv:2004.05150
[3] Kitaev, N., Kaiser, L., and Levskaya, A. (2020). 'Reformer: The efficient transformer'. In International Conference on Learning Representations.
[4] Skowron, M., Trapp, M., Payr, S., and Trappl, R. (2016). 'Automatic identification of character types from film dialogs'. Applied Artificial Intelligence, 30(10):942–973.
[5] Zangerle, E., Mayerl, M., Potthast, M., and Stein, B. (2021). 'Overview of the style change detection task at pan.'