MuLD: The Multitask Long Document Benchmark
G Thomas Hudson, Noura Al Moubayed

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.

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.

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.

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.

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