The NLP Task Effectiveness of Long-Range Transformers

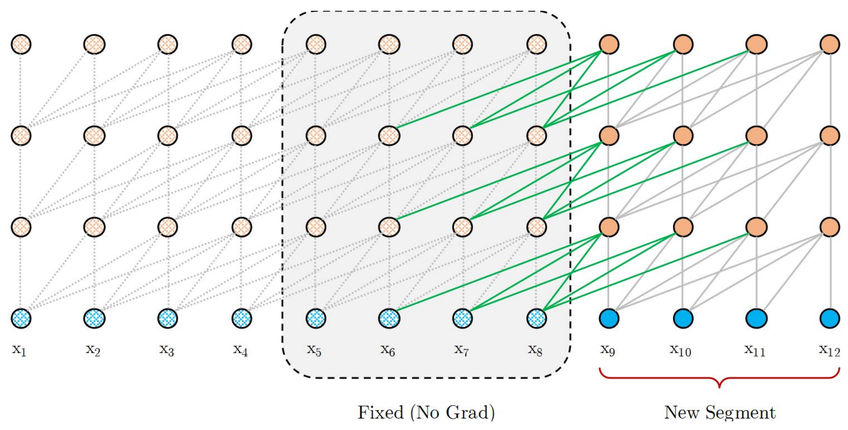
Guanghui Qin, Yukun Feng, and Benjamin Van Durme
Department of Computer Science, Johns Hopkins University

Abstract

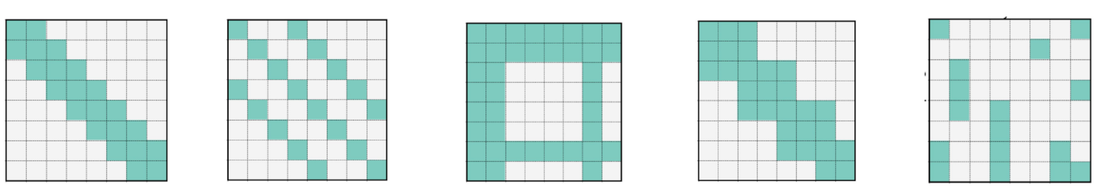
Problem: Evaluate long-range transformers on NLP tasks.
Past work: Simple or non-NLP benchmark.
Experiments: 5 NLP tasks and 7 datasets.
Methods: *Not* a cross-model benchmark. Meant to isolate the effect of pretraining and hyper-parameter settings and to focus on their capacity for long-range attention.
Observations: Advantage & drawbacks of typical long-range models, and the reasons behind their performance.

Types of variants

Recurrence: Attend to past activations



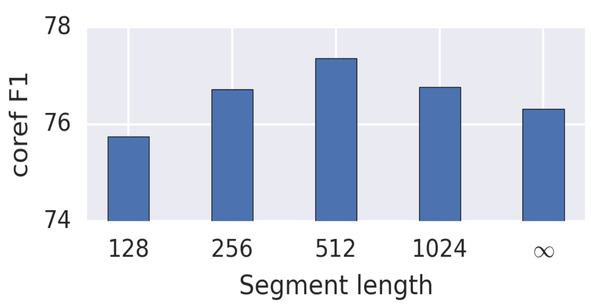
Pattern: Sparsify attention matrix

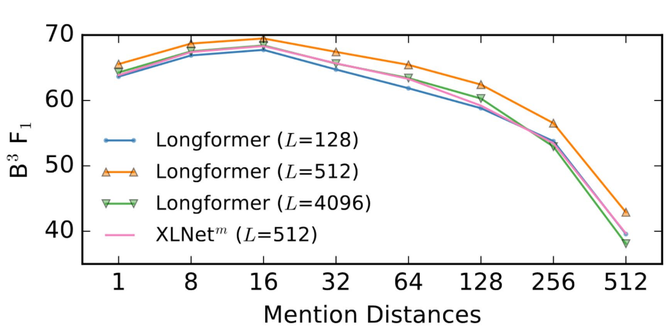


Kernel & low-rank: Approximation

Is long-range attention necessary?

Exp: Longformer on coref
Ablate long-range attention by chunking the texts. Short-range longformer beats the long-range counterpart.

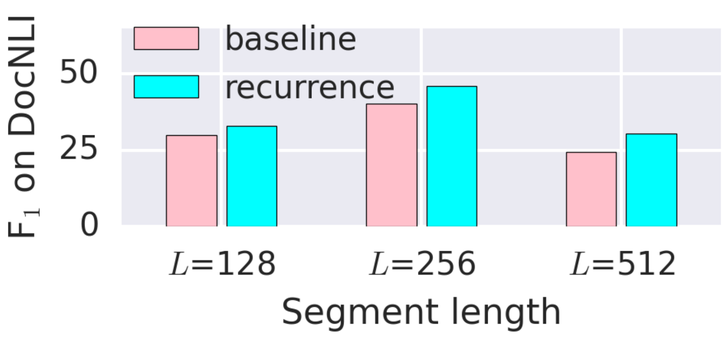


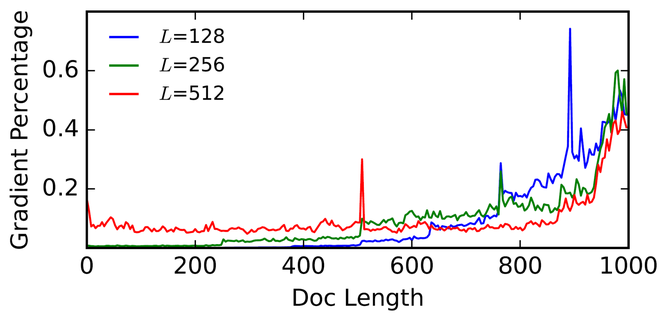


… And it is consistent across different lengths of texts, suggesting that distant information is not exploited by LRT.

Recurrence: Good but can be better

Experiment:
XLNet on DocNLI.
Observation:
Recurrence improve the performance

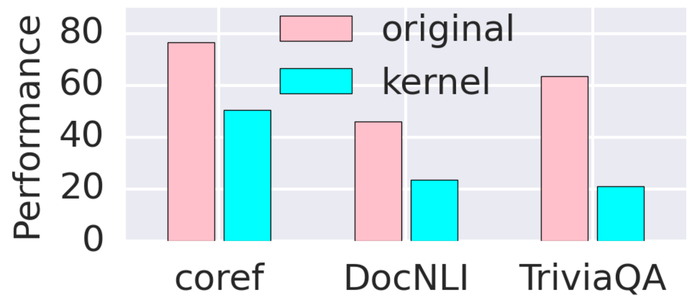


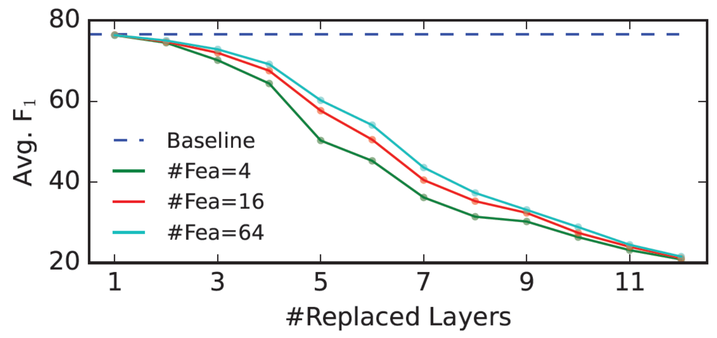


Attribution test:
Contribution of past segments is low. Signals get lost during forward propagation.

Kernel-methods: Unacceptable errors

Replacing BERT attn with kernels leads to dramatic performance drop, possibly due to approxi-mation errors.

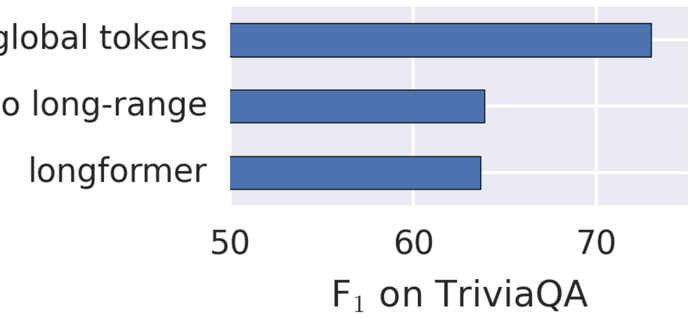


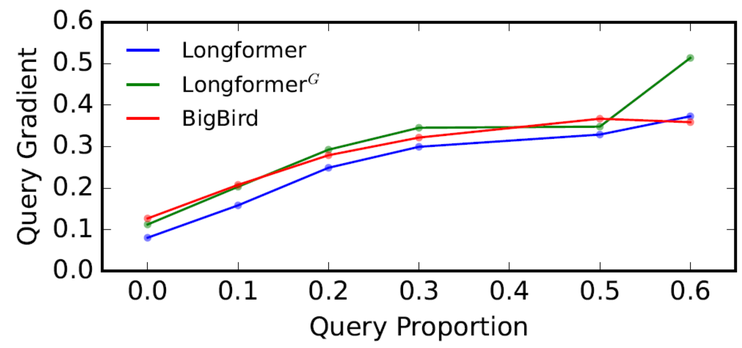


Errors are accumula-ted *layer by layer*, and cannot be fixed with more random features.

Queries as global tokens

Longformer on QA
Boost performance.
Global tokens are the key, not long-range attention.

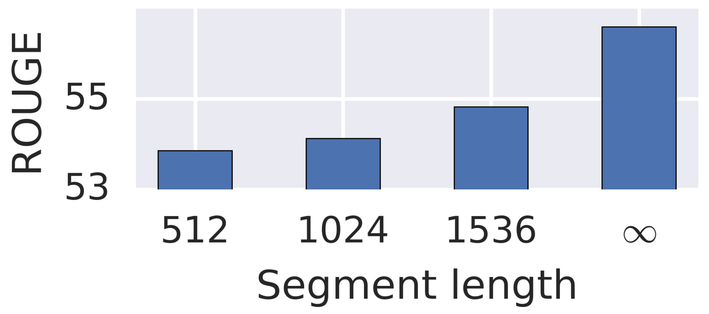


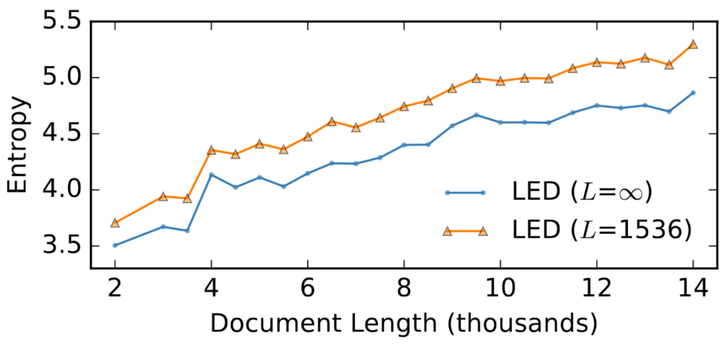


Reason:
Queries are more attended by other tokens when set as global tokens.

Content selection in seq2seq models

Exp: LED on SummFD
Observation:
Long-range attention brings consistent performance boost.





Speculation:
Long-range attn
-> Lower entropy
-> Greater selectivity

More details in the paper! ArXiv: 2202.07856