CANINE:

Pre-training an Efficient Tokenization-Free Encoder

Jon Clark, Dan Garrette, Iulia Turc, John Wieting

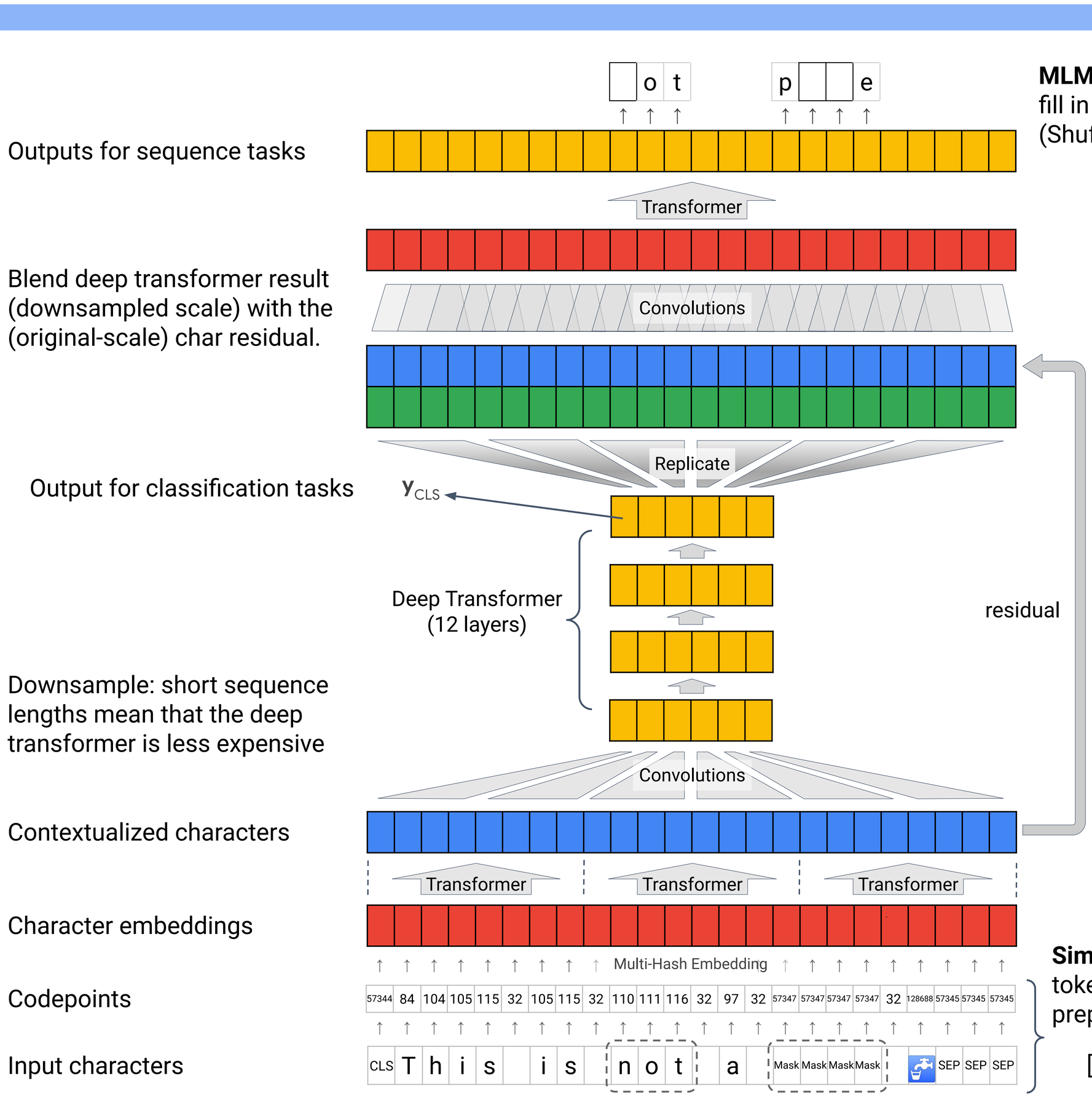
Google Research

Deep neural models have replaced NLP pipelines… except for tokenization.

But language is complex, and tokenizers are brittle.

> Morphological inflection, spelling variation, typos, newly coined terms, domain shift, transliteration, use of digits or punctuation to stand in for letters, languages that don't use whitespace, etc, etc…

Our solution: Let the model operate directly on characters! No tokenizer, no vocab!



MLM Pre-training: Autoregressively fill in the chars of masked spans. (Shuffled order; not left-to-right.)

Simplified engineering: Instead of tokenization models and code, our preprocessing is just one line:

```
[ord(c) for c in text]
```

Model	Input	MLM	Examples /sec	Params	TyDi QA: Passage F1	TyDi QA: MinSpan F1
mBERT (retrained)	Subwords	Subwords	9000	179M	63.2	51.2
	Chars	Single chars	925	127M	59.5 (-3.7)	43.7 (-7.5)
	Chars	Subwords	900	127M	63.8 (+0.6)	50.2 (-1.0)
CANINE-S	Chars	Subwords	6400	127M	66.0 (+2.8)	52.5 (+1.3)
CANINE-C	Chars	Autoregressive chars	6000	127M	65.7 (+2.5)	53.0 (+1.8)
CANINE-C + n-grams	Chars	Autoregressive chars	5600	167M	**68.1** (+4.9)	**57.0** (+5.7)