CANINE

Pre-training an Efficient Tokenization-Free Encoder for Language Representation

Jon Clark, **Dan Garrette**, Iulia Turc, John Wieting

Google Research

Tokenization is Hard

Tokenization is Hard

Spelling variation:

“color” vs “colo**u**r”

Typos / capitalization changes:

```
Queen  Elizabeth  →  Queen     Elizabeth
Queeen elizabeth  →  Que ##een eli ##za ##beth
```

Domain shifts / newly coined terms:

```
COVID-19  →  CO ##VI ##D - 19
```

Tokenization is Hard

Morphological inflection

English **take** → **tak**ing

bet → **bett**ing

Kiswahili **isambaza**yo → **isam** ##**ba** ##**za** ##yo

usambazaji → **usa** ##**mba** ##**zaj** ##i

Arabic **k**-**t**-**b** → **k**a**t**a**b**a

Finnish saap**ua** → saav**uin**

jumit**tua** → jumitu**in**

Tokenization is Hard

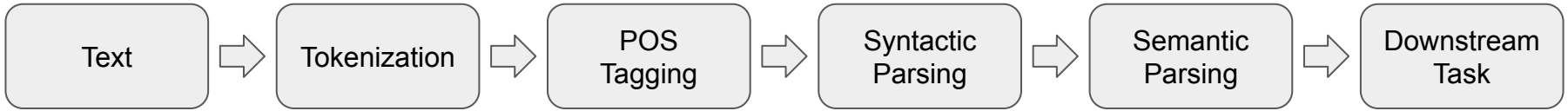
Some languages don’t use whitespace:

Chinese, Japanese, Thai, Khmer, Lao, Burmese, …

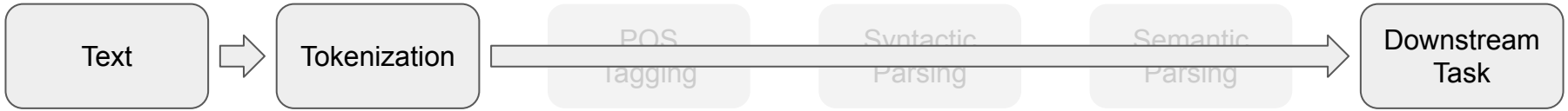
Vietnamese is typically written with spaces between *syllables*.

Tokenization is Hard: Skip it!

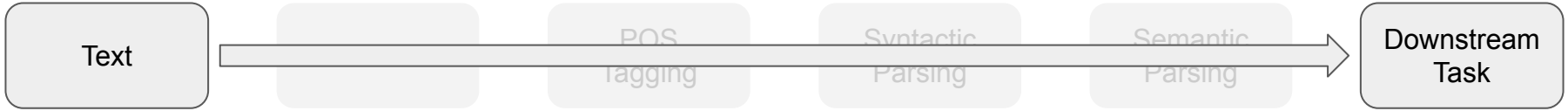
Classic NLP pipeline



Current standard



Our approach



Token-Free Approach

Token-Free Approach

No Tokenizer

- Operate directly on characters.

No Vocabulary

- Full Unicode codespace ($0–10FFFF_{hex}$)
  - All 1.1M current *and future* Unicode characters. (No OOV.)

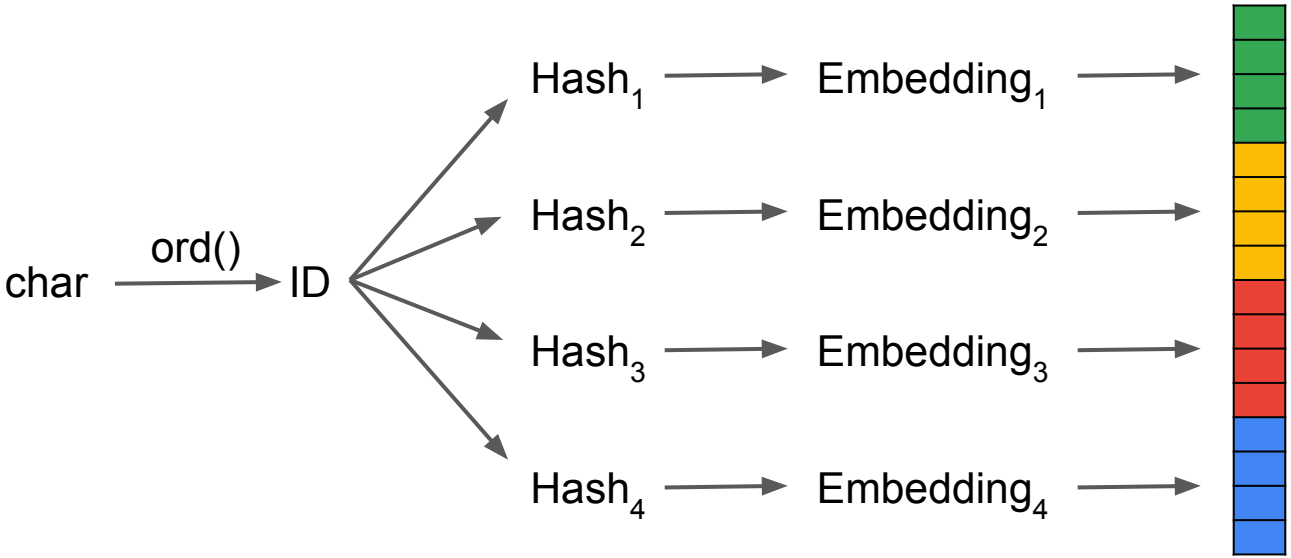
Token-Free Approach

Preprocessing implementation (Python):

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

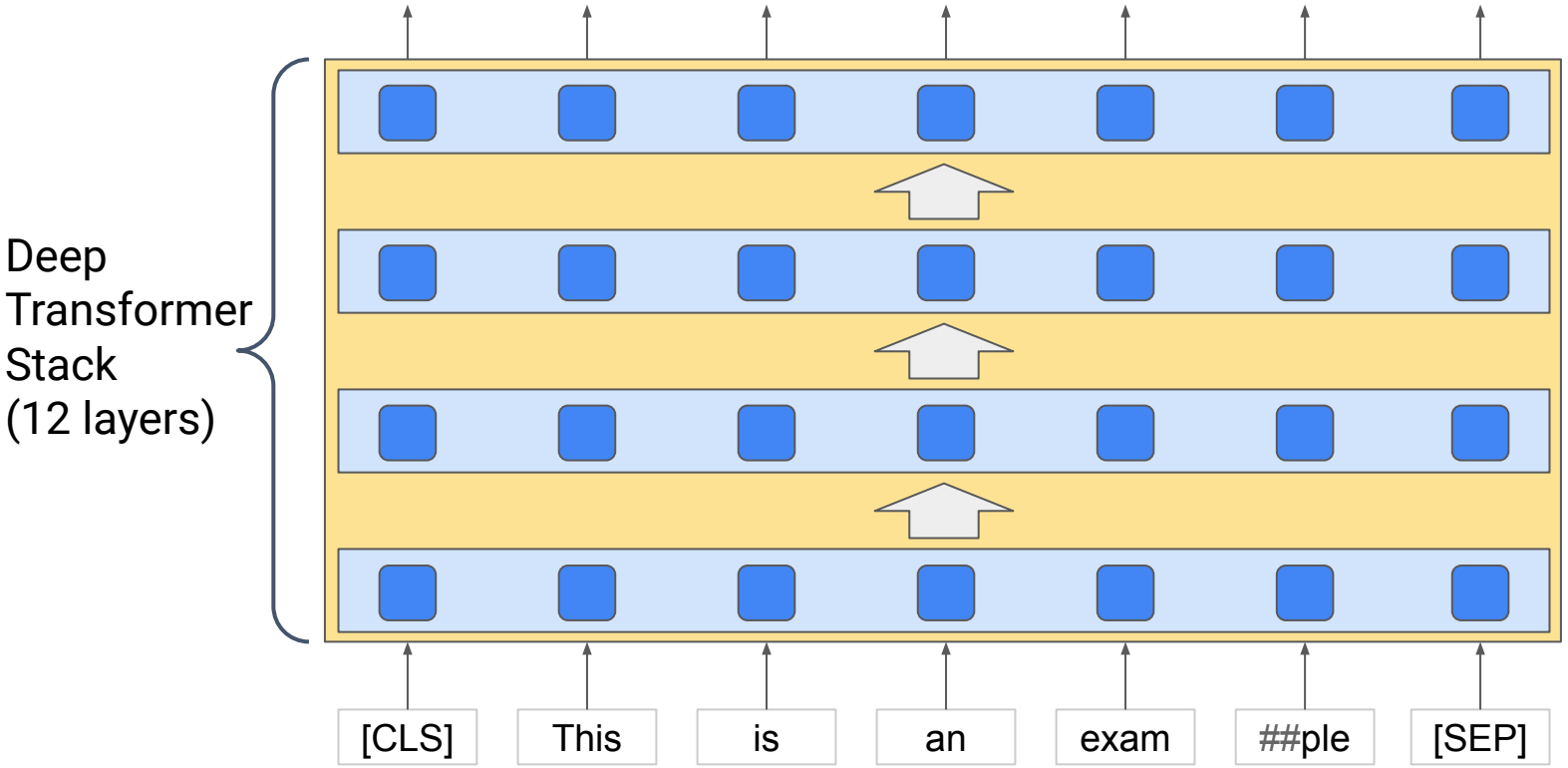
Multi-Hash Embedding

How to embed all 1.1M codepoint values?

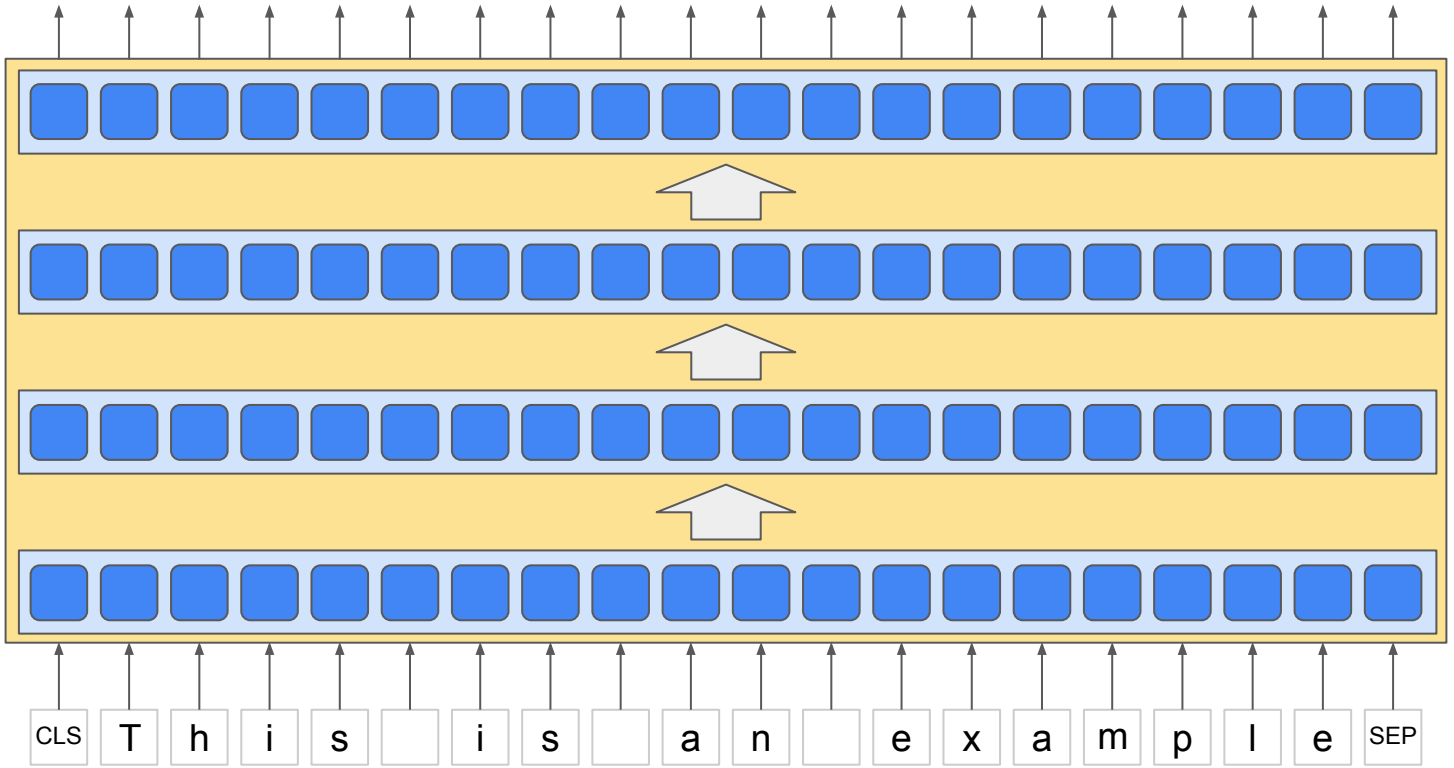


Model

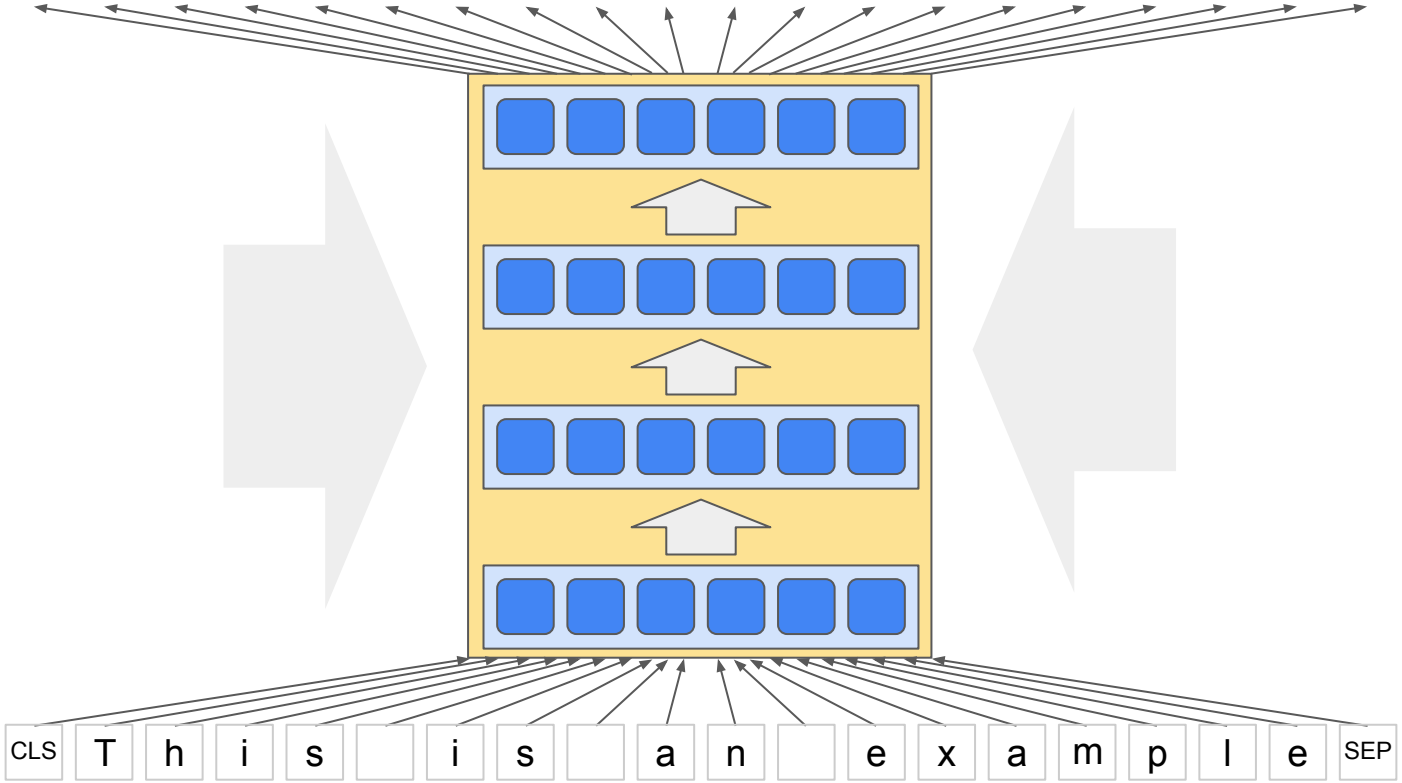
Baseline: BERT



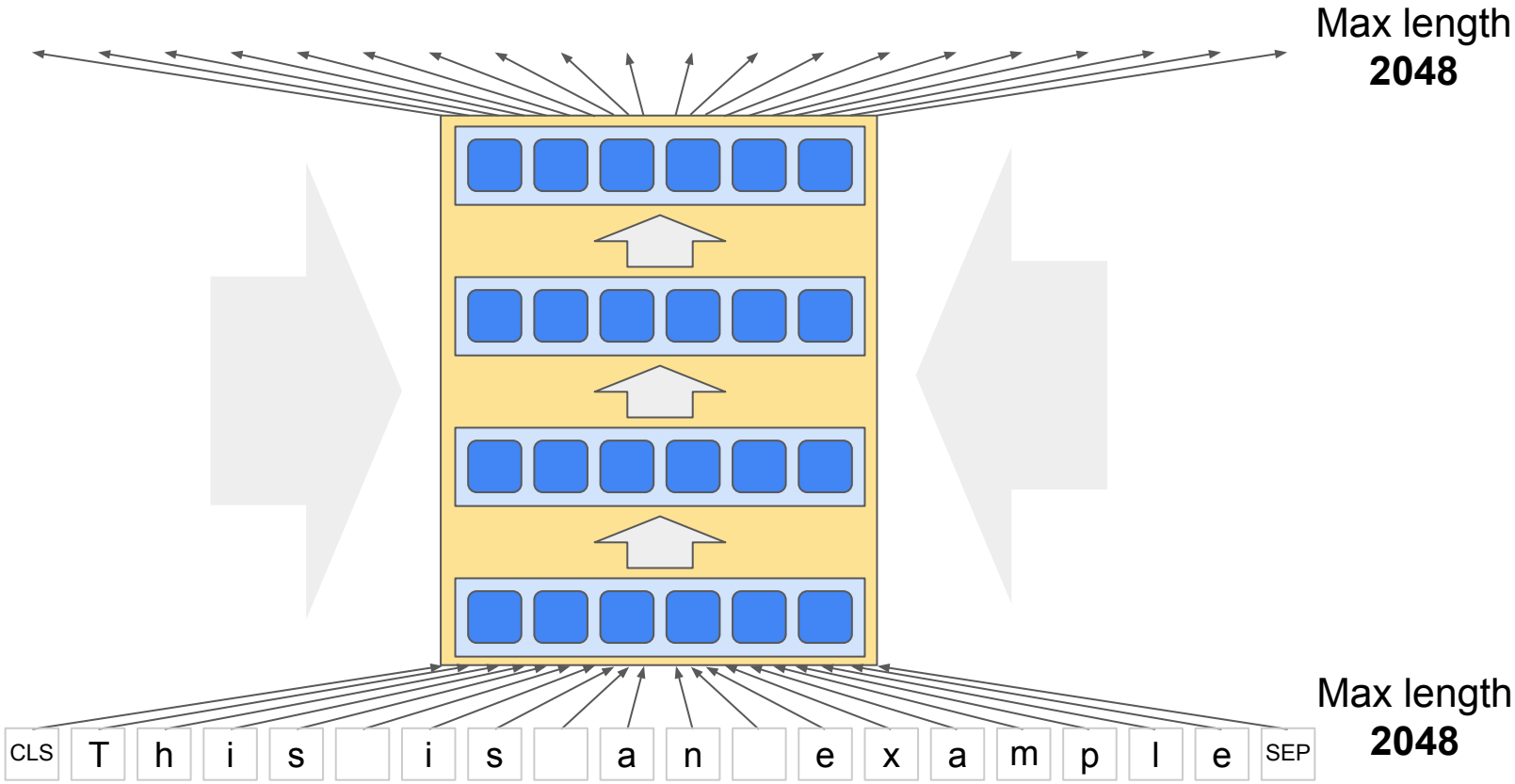
Baseline: BERT, but characters (10x slower)



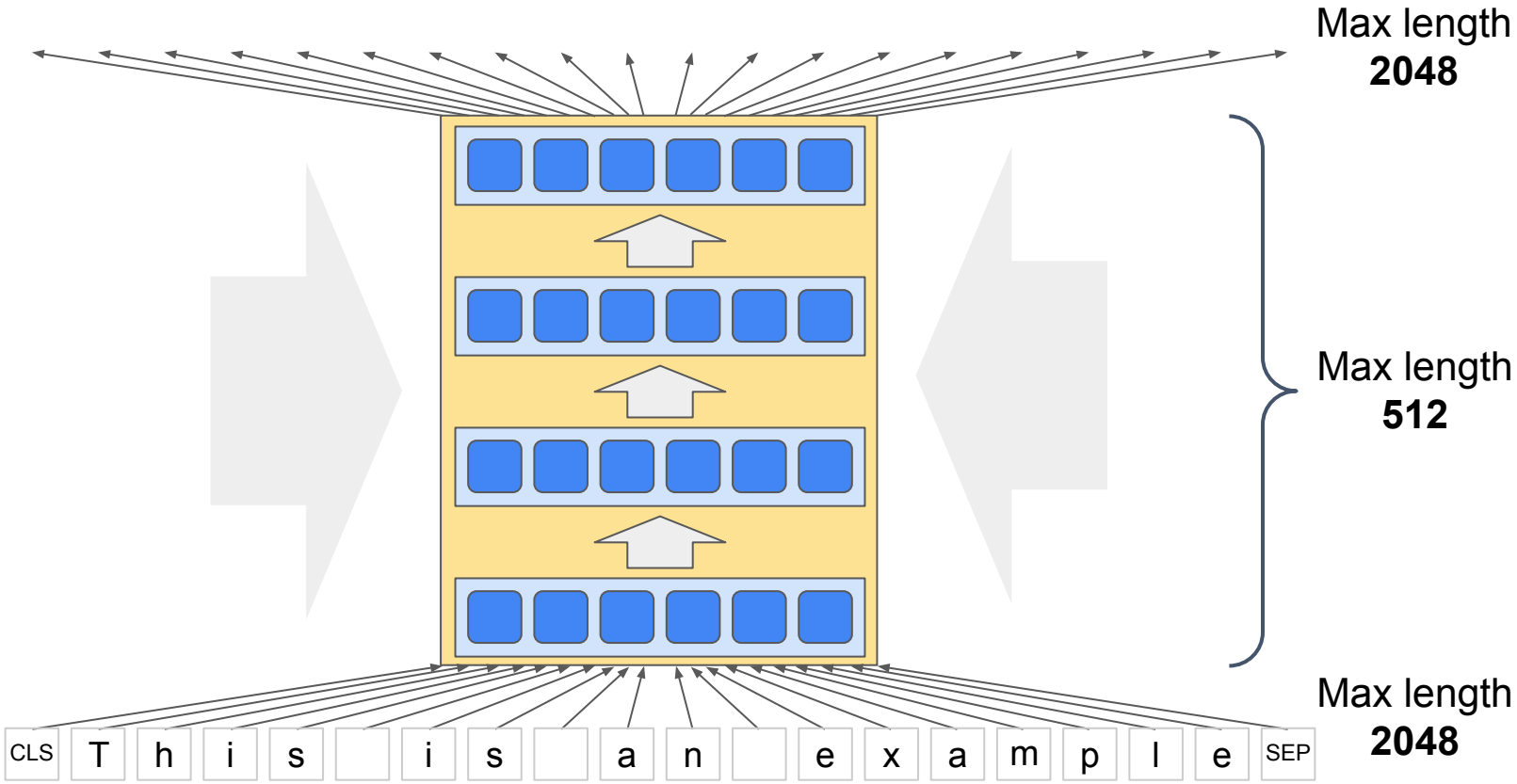
What We Want



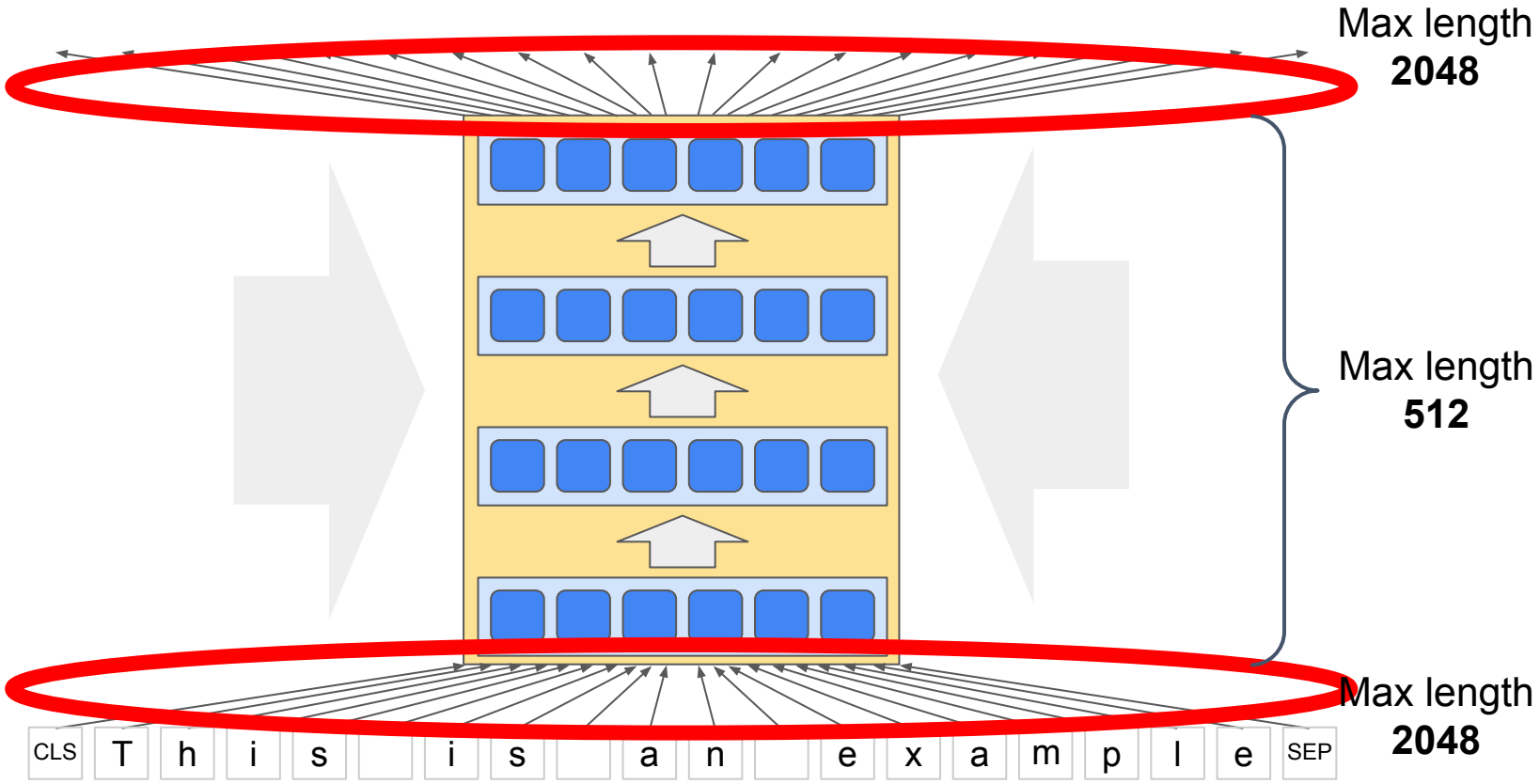
What We Want



What We Want

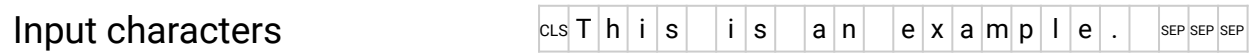

Max length
2048
Max length
512
Max length
2048
CLS
T
h
i
s
i
s
a
n
e
x
a
m
p
l
e
SEP

What We Want


Max length
2048
Max length
512
Max length
2048
CLS
T
h
i
s
i
s
a
n
e
x
a
m
p
l
e
SEP

CANINE

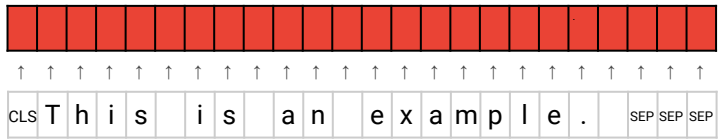
CANINE



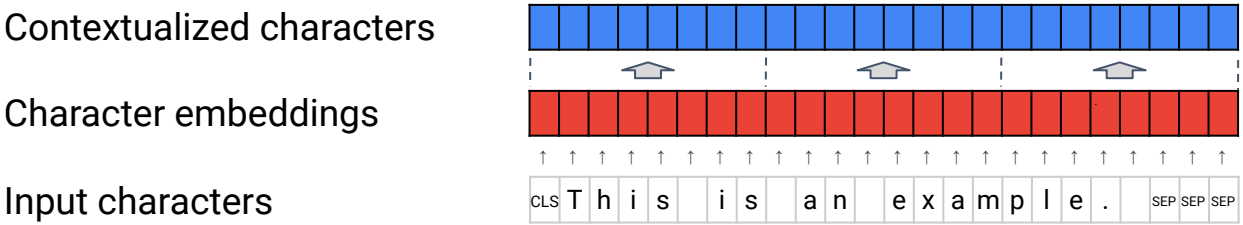
CANINE

Character embeddings

Input characters



CANINE


Contextualized characters
Character embeddings
Input characters
CLS T h i s i s a n e x a m p l e . SEP SEP SEP

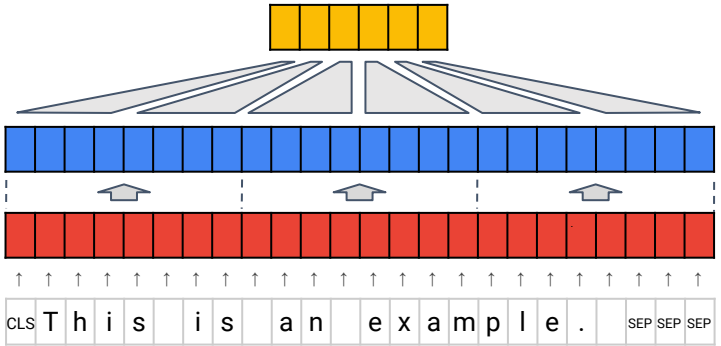
CANINE

Downsampled

Contextualized characters

Character embeddings

Input characters



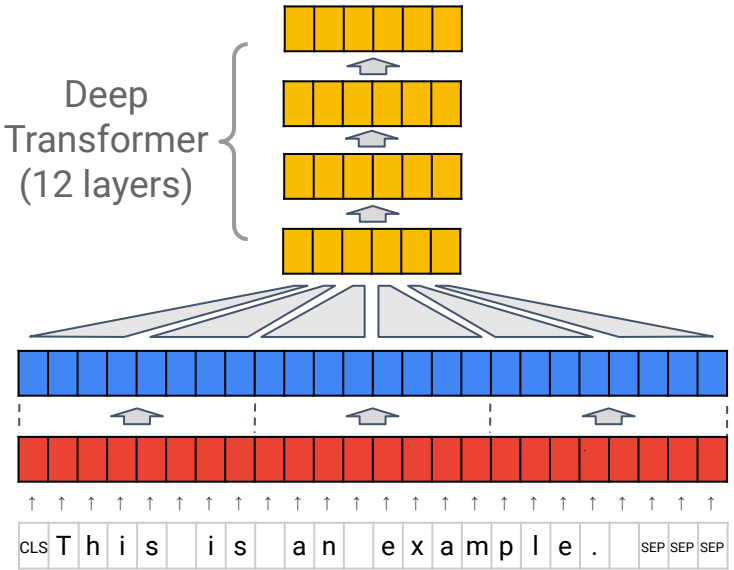
CANINE

Downsampled

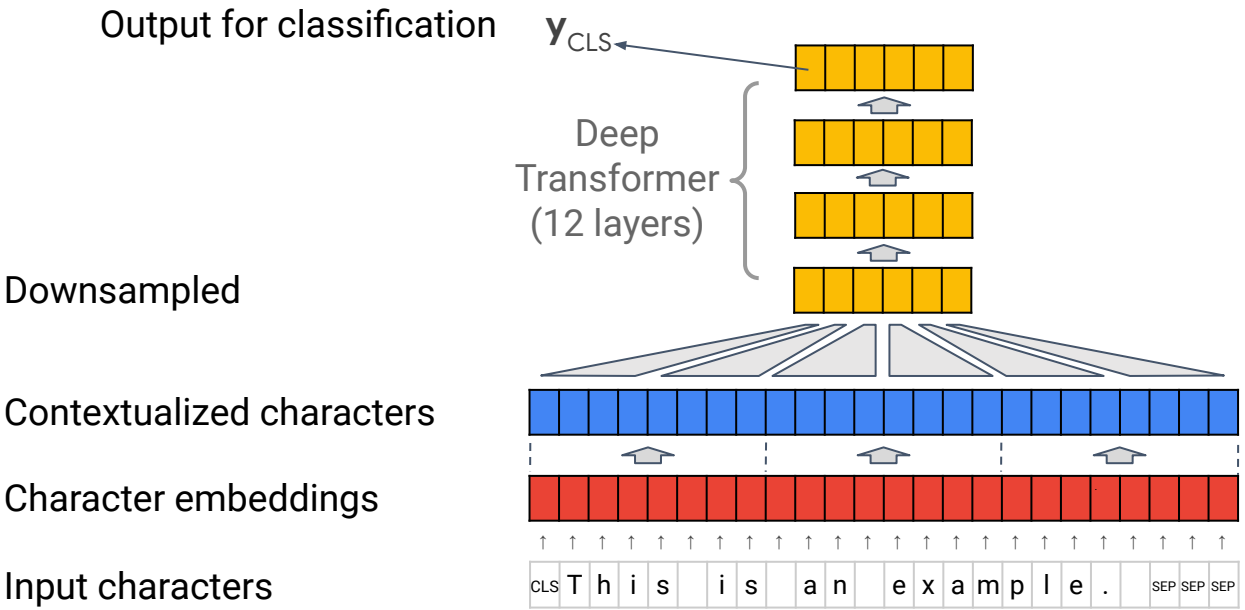
Contextualized characters

Character embeddings

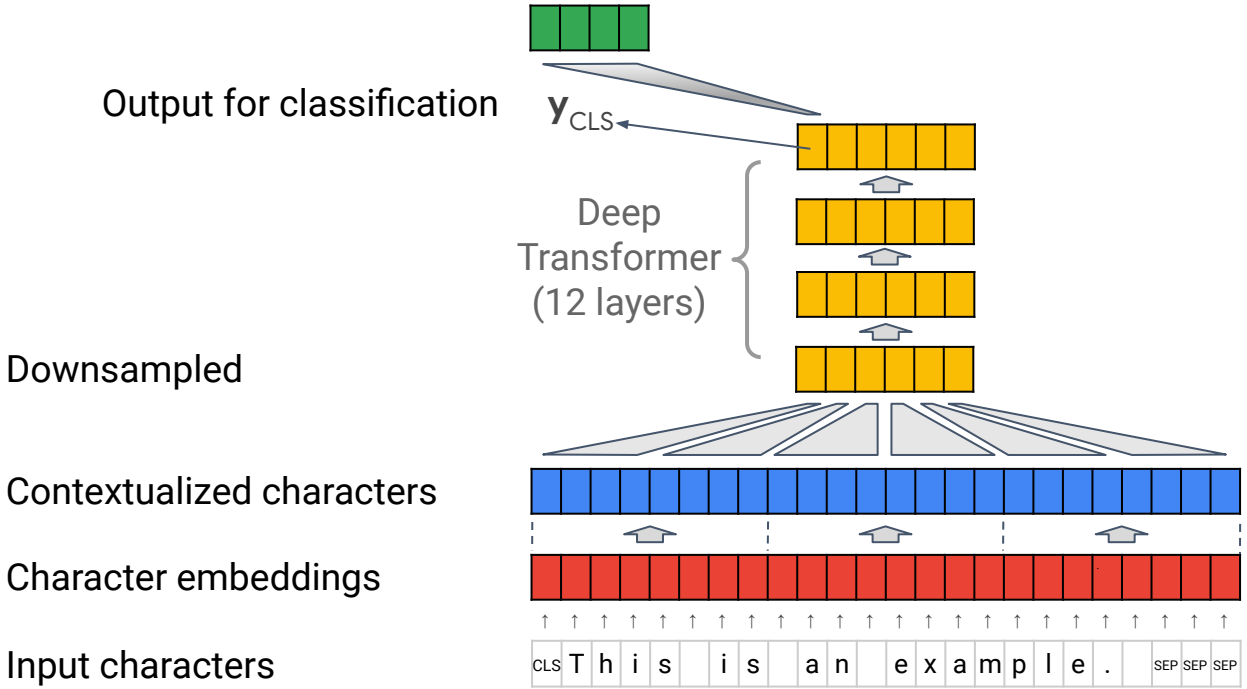
Input characters



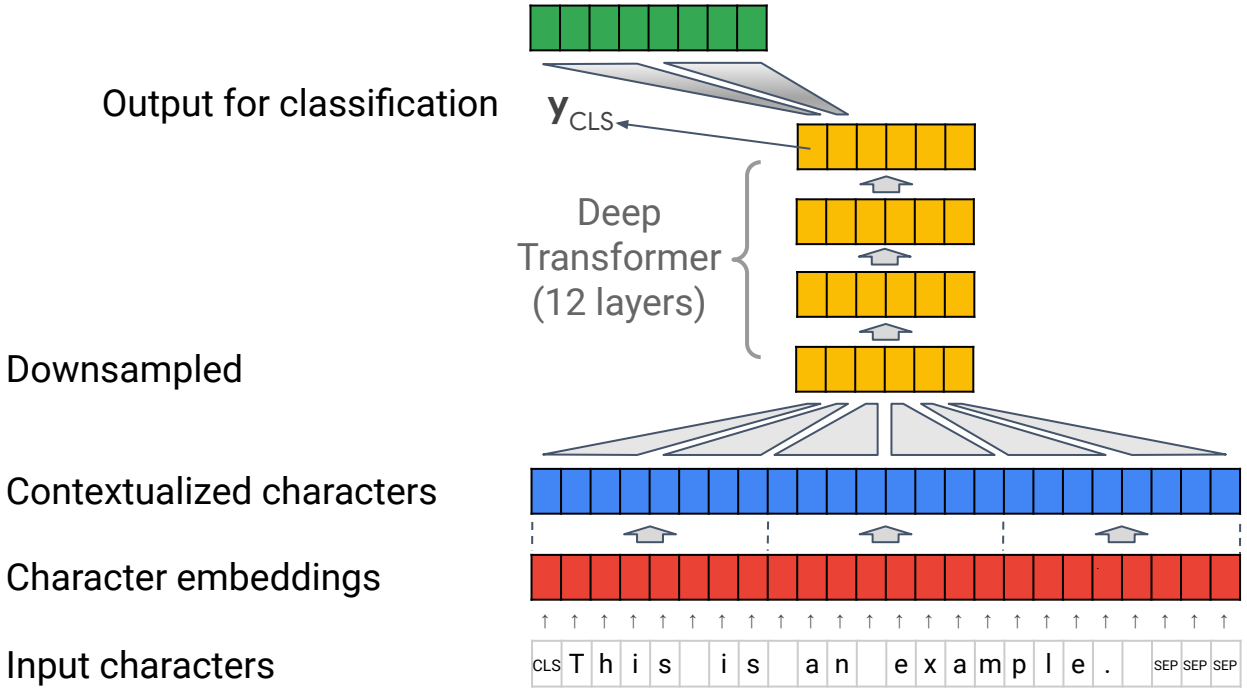
CANINE


Output for classification
y_{CLS}
Deep Transformer (12 layers)
Downsampled
Contextualized characters
Character embeddings
Input characters
CLS This is an example. SEP SEP SEP

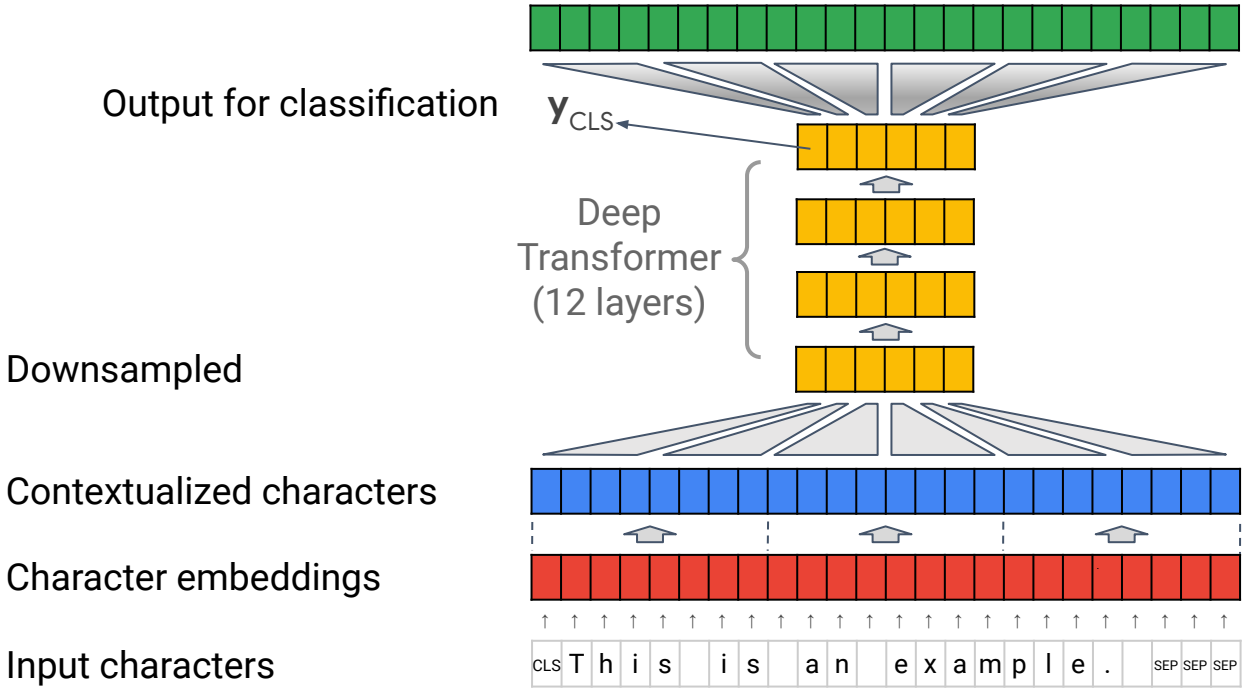
CANINE


Output for classification
y_{CLS}
Deep Transformer (12 layers)
Downsampled
Contextualized characters
Character embeddings
Input characters
CLS This is an example. SEP SEP SEP

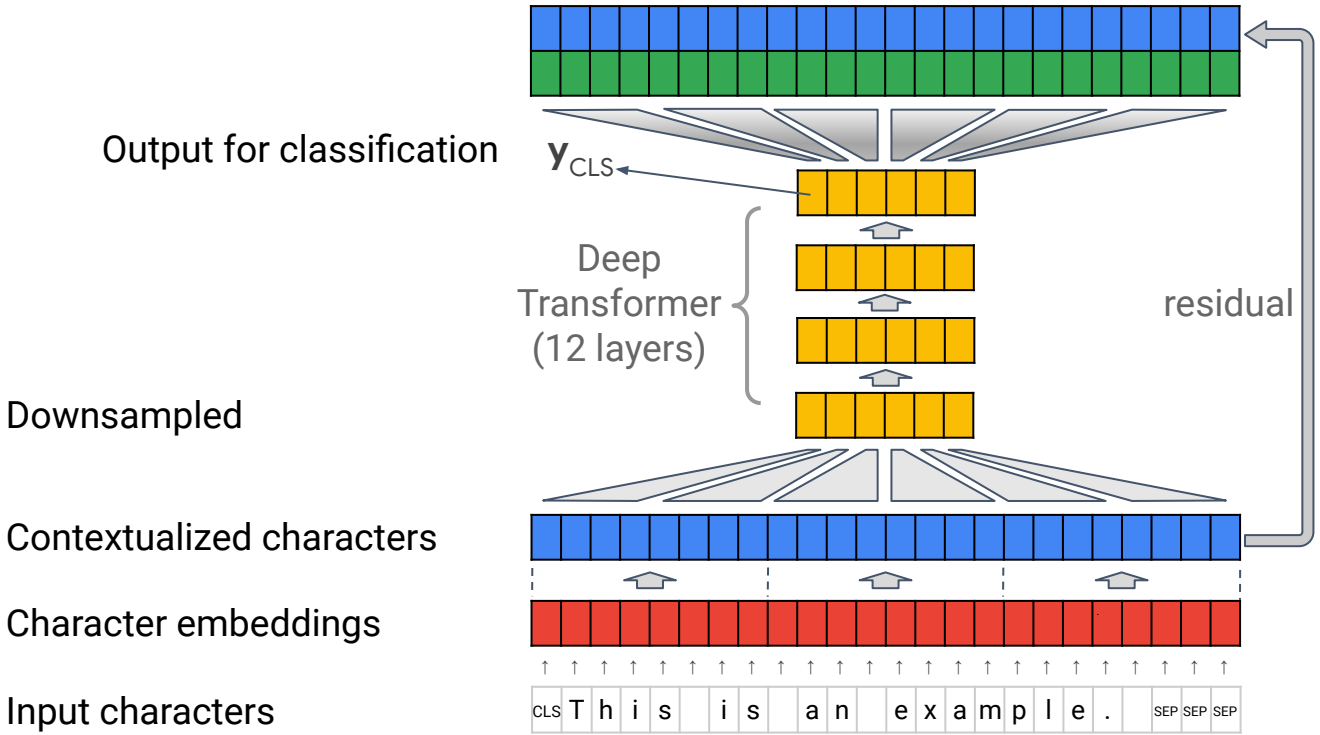
CANINE


Output for classification
y_{CLS}
Deep Transformer (12 layers)
Downsampled
Contextualized characters
Character embeddings
Input characters
CLS This is an example. SEP SEP SEP

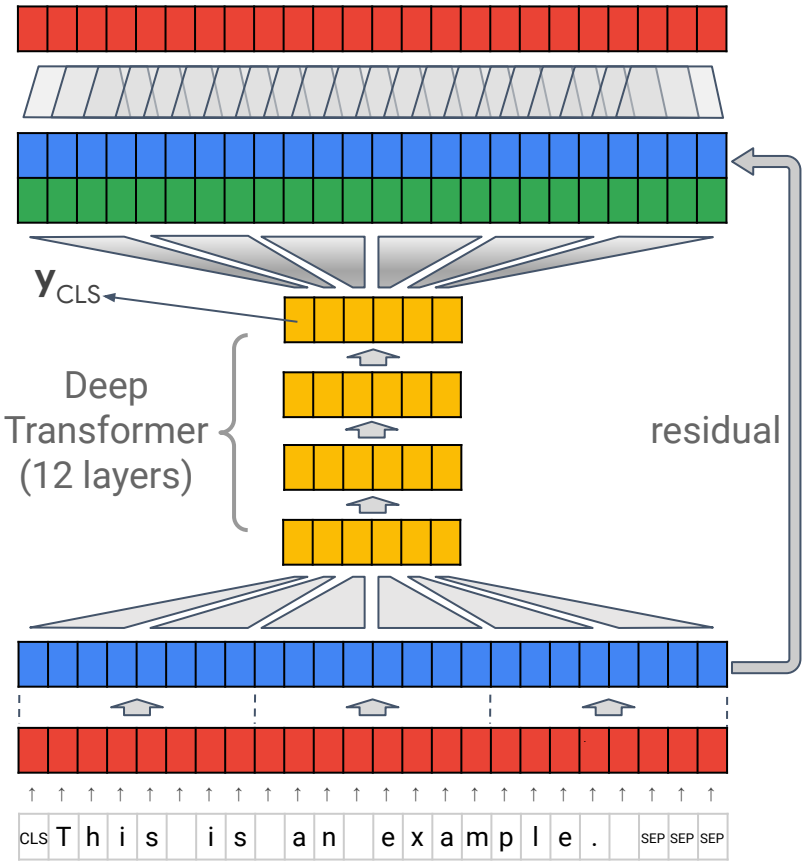
CANINE


Output for classification
y_{CLS}
Deep Transformer (12 layers)
Downsampled
Contextualized characters
Character embeddings
Input characters
CLS This is an example. SEP SEP SEP

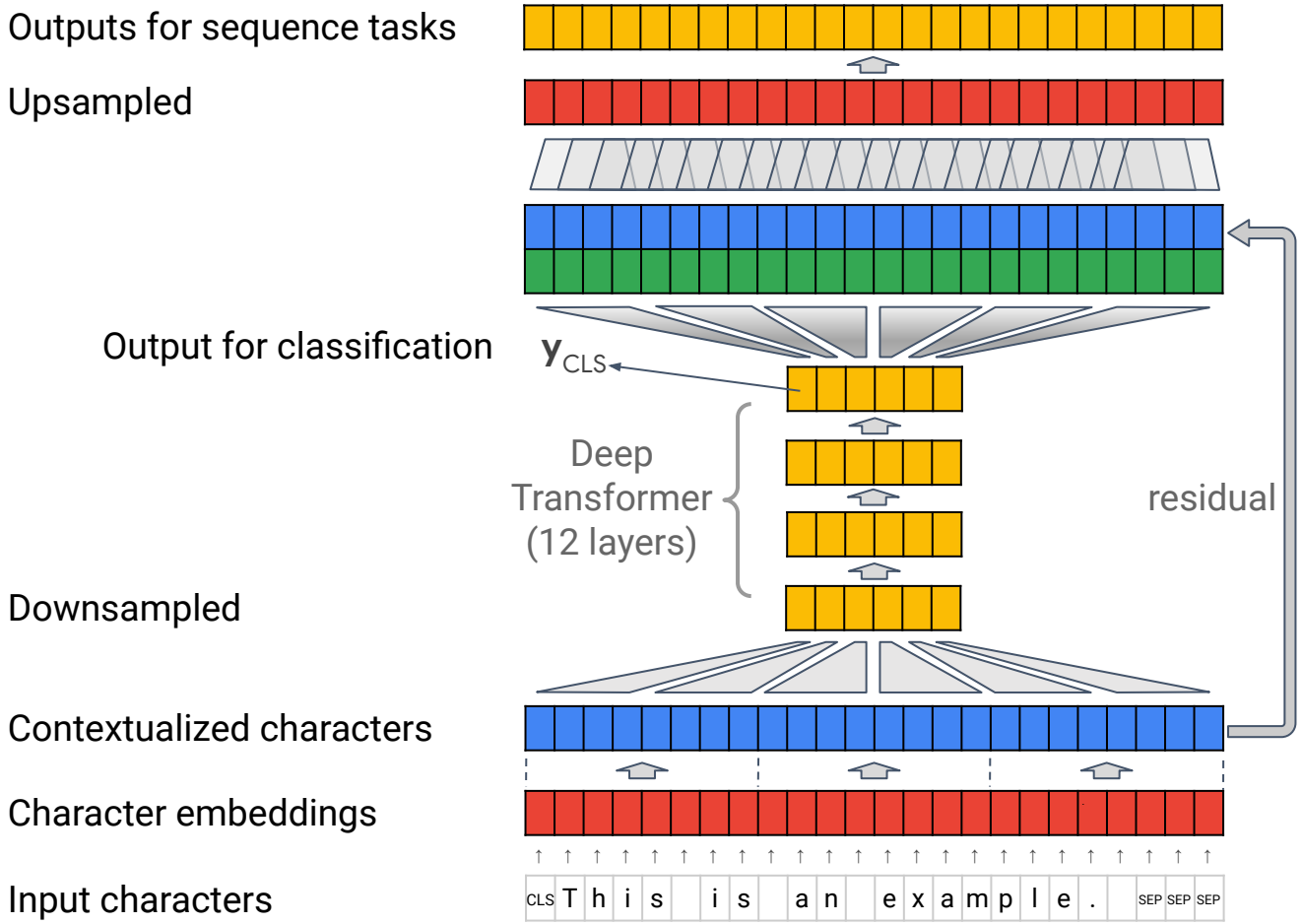
CANINE


Output for classification
y_{CLS}
Deep Transformer (12 layers)
residual
Downsampled
Contextualized characters
Character embeddings
Input characters
CLS This is an example. SEP SEP SEP

CANINE


Upsampled
Output for classification
y_{CLS}
Deep Transformer (12 layers)
residual
Downsampled
Contextualized characters
Character embeddings
Input characters
CLS This is an example. SEP SEP SEP

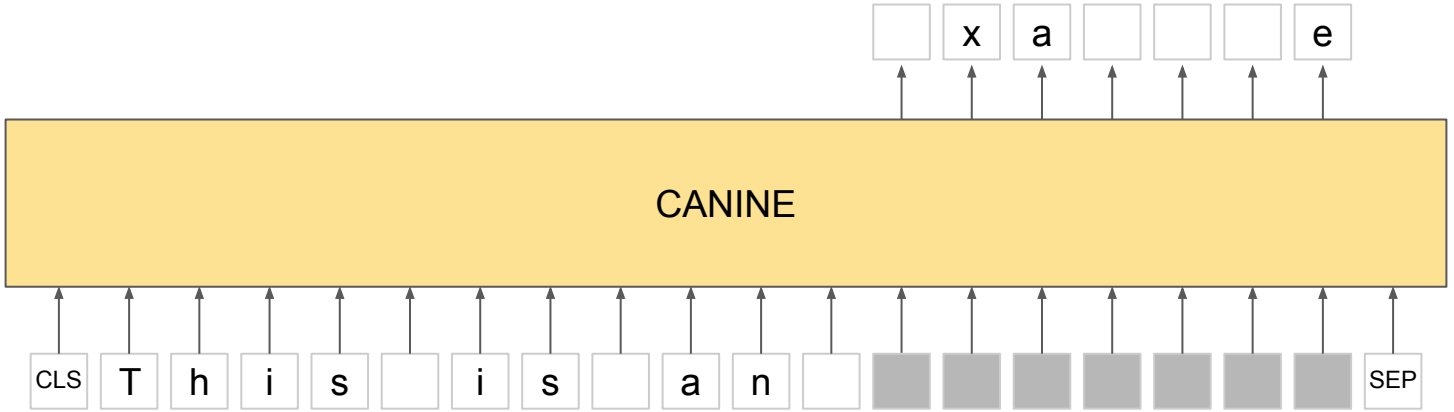
CANINE


Outputs for sequence tasks
Upsampled
Output for classification
y_{CLS}
Deep Transformer (12 layers)
residual
Downsampled
Contextualized characters
Character embeddings
Input characters
CLS This is an example. SEP SEP SEP

Pre-training

MLM Pre-training

Auto-regressively predict each masked character (shuffled order, not left-to-right).



Experimental Results

Experimental Results

Model	Input	MLM	Examples /sec	Params	TyDi QA: Passage F1	TyDi QA: MinSpan F1
mBERT (retrained)	Subwords	Subwords	9000	179M	63.2	51.2

Experimental Results

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)

Experimental Results

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)

Experimental Results

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)

Experimental Results

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	Auto-reg. chars	6000	127M	65.7 (+2.5)	53.0 (+1.8)

Experimental Results

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	Auto-reg. chars	6000	127M	65.7 (+2.5)	53.0 (+1.8)
CANINE-C + n-grams	Chars	Auto-reg. chars	5600	167M	**68.1** (+4.9)	**57.0** (+5.7)

Experimental Results

	TyDi QA: Passage F1	TyDi QA: MinSpan F1
(English)	+2.4	+5.8
Arabic	+2.0	+2.3
Bengali	+7.5	+9.8
Finnish	+6.3	+6.0
Indonesian	+4.6	+4.6
Japanese	+5.0	+5.9
Korean	+0.4	+3.1
Russian	+6.3	+5.9
Swahili	+8.4	+9.8
Telugu	+3.6	+4.1
Thai	+4.7	+5.8
Macro Avg	+4.9	+5.7

Conclusion

Conclusion

- CANINE: Tokenization-free encoder.
  - Operates directly on input **characters**.
  - **Higher quality** than comparable subword-based model across a **variety of languages**.
  - Downsampling architecture **mitigates slowdown** from increased sequence length.
- Models and code available for download, and in HuggingFace Transformers.
- On-going work with ByT5 authors to explore new token-free approaches.