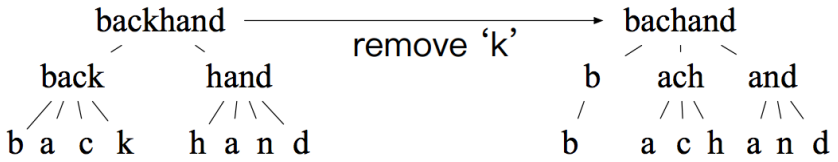
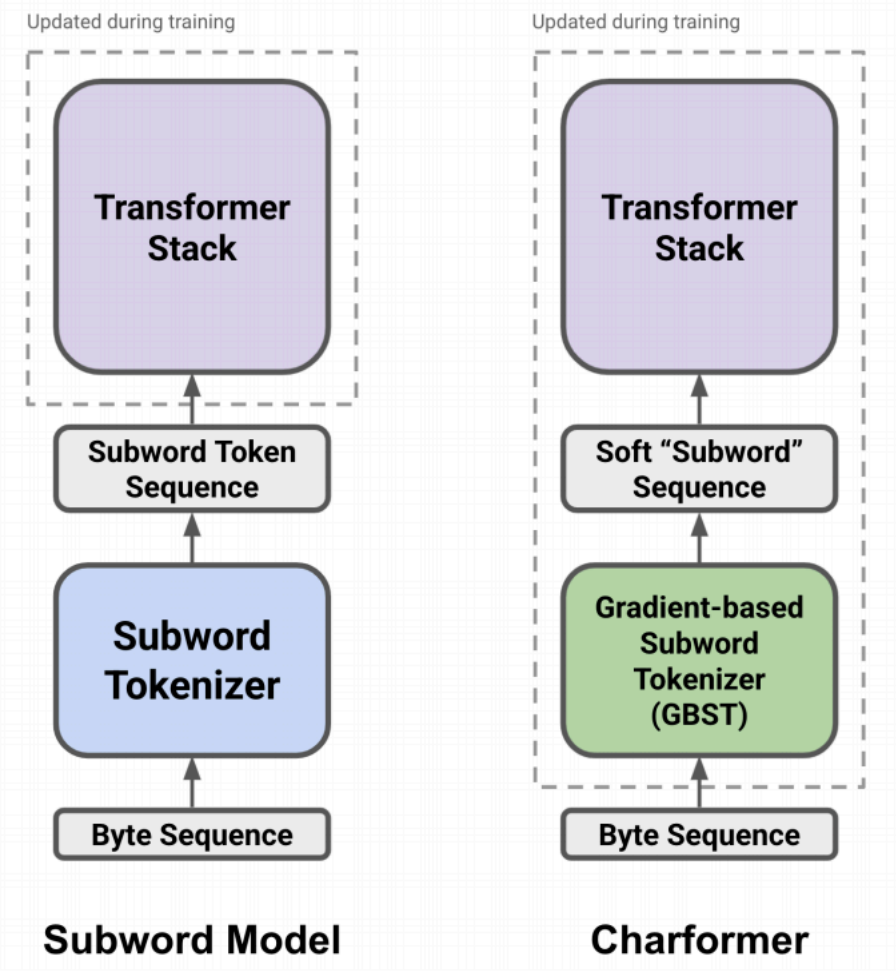
Charformer: Fast Character Transformers via Gradient-based Subword Tokenization

Tay et al., arXiv 21

Overview

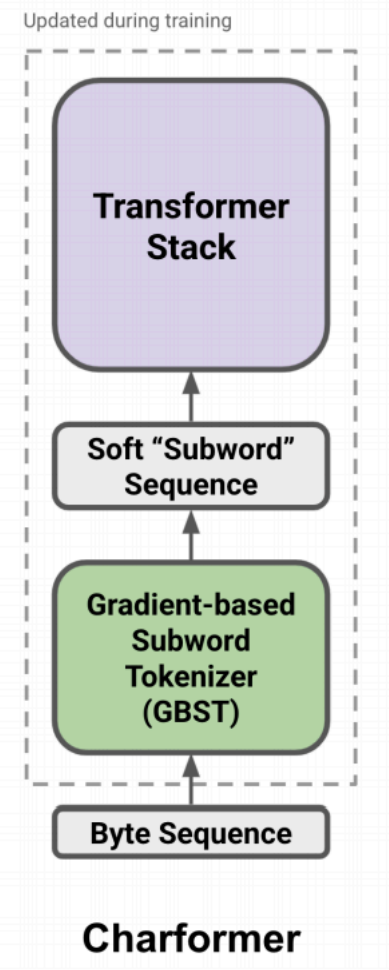
- Trend: Subword-based tokenization
  - Brittle to rare words, perturbations
  - Impact on low-resource languages
  - Mismatch in pre-training and downstream distribution of words
- Quick solution: Char level models
  - Compute/memory complexity
  - Lower performance
- Proposal: Learn latent subword representation from characters





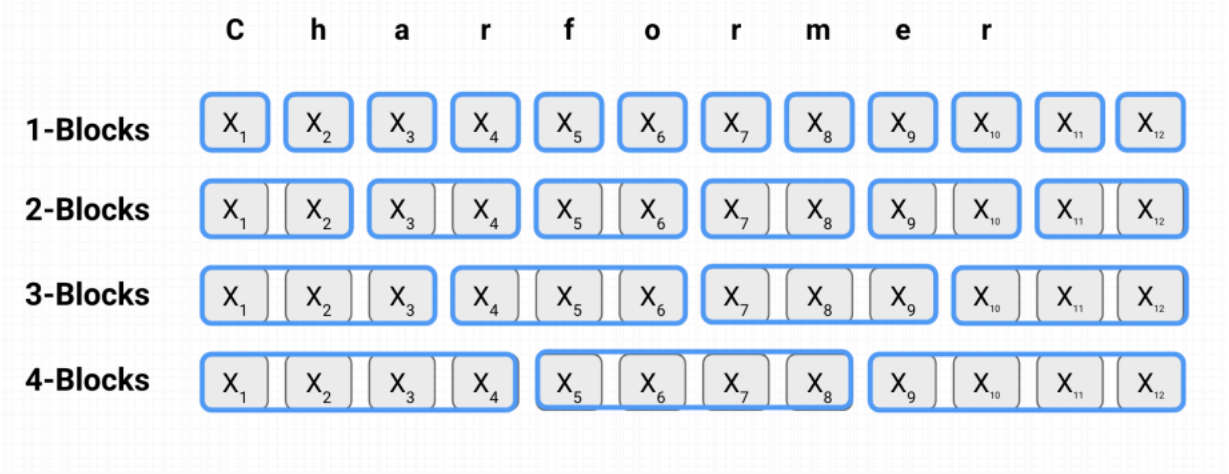
Charformers

- Gradient-based subword tokenization (GBST)
  - Learn latent subword representations from characters
    - Learns a position-wise soft selection over candidate subword blocks by scoring them with a scoring network
  - Interpretable latent subword
- Charformer - Encoder-Decoder model (byte level)
- English (standard/non-standard), Multilingual eval
- Efficiency (Memory, speed)



Charformer

- Gradient-based subword tokenization
  - Constructing Candidate Latent Subword Blocks
  - Block Scoring Network
  - Forming Latent Subwords
  - Position-wise Score Calibration
  - Downsampling
- Transformer stack

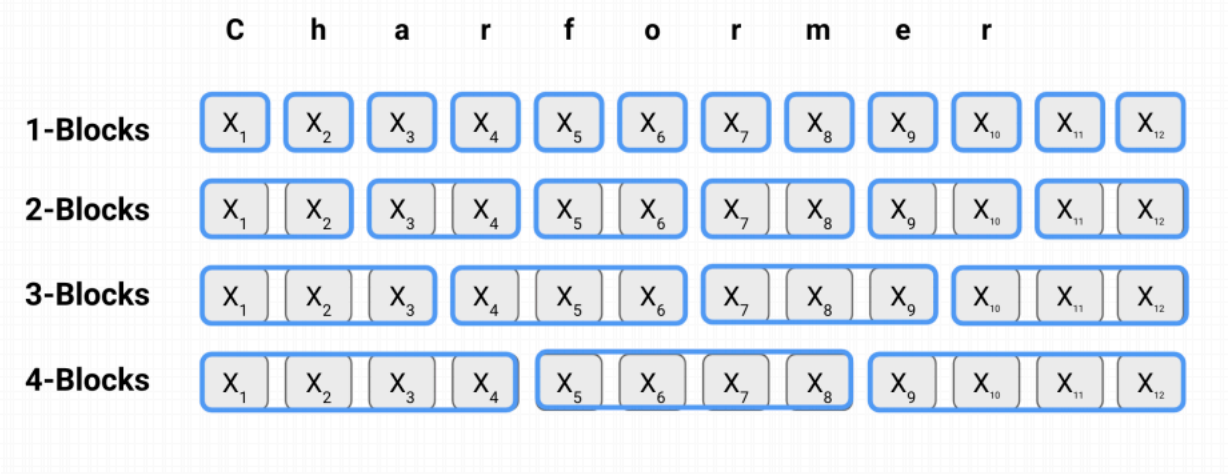


Charformer

- Gradient-based subword tokenization
  - Input: $X \in \mathbb{R}^{L \times d}$
  - Block = Contiguous span of characters $X_{i:i+b}$
  - Constructing Candidate Latent Subword Blocks

$$X_b = [F(X_{i:i+b}); F(X_{(i+s):(i+s)+b}); \ldots]$$

    - Offsets – ‘h’, ‘a’ missed when b=2
      - Slide window vs. ConvNet output
    - Intra-block positions – intra block position
      - Position embeddings vs. ConvNet mean pooling



Charformer

- Gradient-based subword tokenization
  - Block Scoring Network

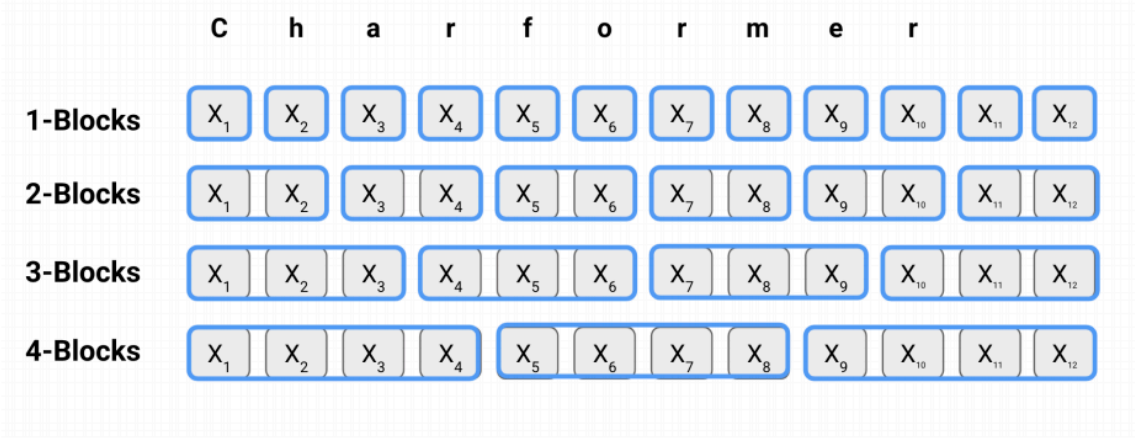
$$p_{b,i} = F_R(X_{b,i})$$

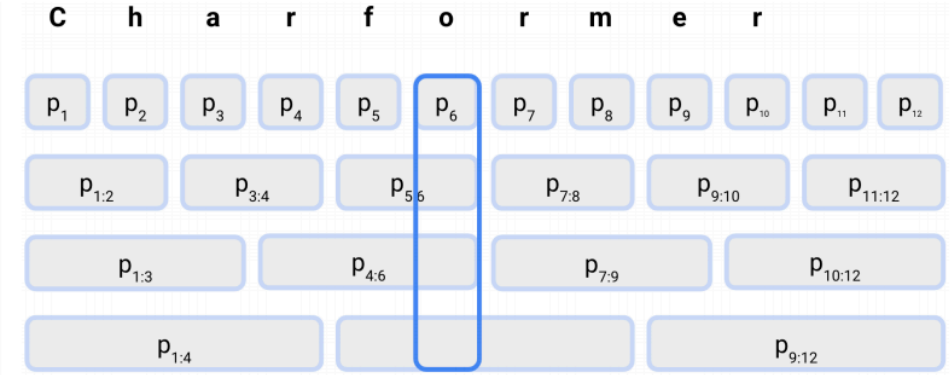
$$P_i = \text{softmax}([p_{0,i}, p_{1,i}, \cdots, p_{M,i}])$$

  - Forming latent subwords

$$\hat{X}_i = \sum_{b}^{M} P_{b,i} X_{b,i}$$

  - Position-wise calibration $\hat{P} = \text{softmax}(PP^\top)P.$





Charformer

- Downsampling
  - Mean pooling with stride $F_D : \mathbb{R}^{L \times d} \rightarrow \mathbb{R}^{\frac{L}{d_s} \times d}$
- Transformer stack
  - T5 pretraining scheme
  - Tall model – Narrow, deep encoder + small decoder
  - Compare against on char level model on params, compute and inference cost

Evaluation - Standard English Datasets

Table 1: Comparison of CHARFORMER against other subword and character-level models with different parameter sizes on diverse standard English datasets.

Model	$\|\theta\|$	SST-2	MNLI	QNLI	MRPC	QQP	STSB	COLA
BERT$_{Base,Subword}$	110M	92.7	84.4/-	88.4	86.7/-	-	-	-
T5$_{Base,Subword}$	220M	92.7	84.2/84.6	90.5	88.9/92.1	91.6/88.7	88.0	53.8
Byte-level T5$_{Small}$	45M	89.2	**79.7**/79.9	86.7	83.6/88.5	90.2/86.7	82.1	27.3
Funnel T5$_{Small}$	45M	89.6	78.7/79.2	86.4	84.8/89.5	90.2/86.8	**84.1**	28.8
Byte-level T5+LASC$_{Small}$	47M	87.5	75.6/76.1	84.2	80.4/86.3	88.1/84.2	81.1	17.3
CHARFORMER$_{Small}$	48M	**90.4**	79.2/**80.2**	**87.3**	**84.8/89.4**	**90.4/87.1**	83.6	**32.4**
Byte-level T5$_{Base}$	200M	**91.6**	82.5/**82.7**	88.7	**87.3**/91.0	90.9/87.7	84.3	**45.1**
Byte-level T5+LASC$_{Base}$	205M	90.0	80.0/80.8	87.1	82.8/88.1	89.0/85.4	83.7	25.3
CHARFORMER$_{Base}$	203M	**91.6**	**82.6/82.7**	**89.0**	**87.3/91.1**	**91.2/88.1**	**85.3**	42.6
CHARFORMER$_{Tall}$	134M	91.5	**83.7/84.4**	**91.0**	**87.5/91.4**	**91.4/88.5**	87.3	51.8

Evaluation – Noisy & Long Documents

Table 2: Results on comment classification on Civil Comments and Wiki Comments. Metrics are accuracy and AUC-PR. T5 baseline results are from [Tay et al., 2021].

Model	Civil Comments	Wiki Comments
$\text{T5}_{Base,Subword}$	81.2 / -	91.5 / -
Byte-level T5_{Small}	83.1 / 78.6	92.9 / 76.2
Byte-level T5+LASC_{Small}	82.4 / 77.2	92.9 / 76.3
$\text{CHARFORMER}_{Small}$	**83.1 / 78.7**	**93.3 / 78.2**
Byte-level T5_{Base}	82.8 / 78.7	93.2 / 75.4
Byte-level T5+LASC_{Base}	82.9 / 78.2	93.0 / 75.0
CHARFORMER_{Base}	**83.0** / 78.8	92.7 / **79.7**
CHARFORMER_{Tall}	**83.0 / 78.9**	**93.5** / 75.5

Table 3: Results on text classification on long documents.

Model	IMDb	News
$\text{T5}_{Base,Subword}$	94.2	93.5
Byte-level T5_{Small}	90.1	93.8
Funnel T5_{Small}	90.6	93.5
Byte-level T5+LASC_{Small}	90.6	92.5
$\text{CHARFORMER}_{Small}$	90.6	**93.9**
Byte-level T5_{Base}	91.5	93.6
Byte-level T5+LASC_{Base}	91.1	93.5
CHARFORMER_{Base}	91.5	94.0
CHARFORMER_{Tall}	**94.4**	**94.1**

Evaluation – Multilingual Documents

		In-Language	Translate-Train-All				Zero-Shot	
Model	$\|\theta\|$	TyDiQA-GoldP	XQuAD	MLQA	XNLI	PAWS-X	XNLI	PAWS-X
mBERT$_{Base}$ (Subword)	179M	77.6/68.0	-/-	-/-	-	-	65.4	81.9
mT5$_{Base}$ (Subword)	582M	80.8/70.0	75.3/59.7	67.6/48.5	75.9	89.3	75.4	86.4
Byte-level T5$_{Small}$	45M	**71.6/60.7**	**64.6/50.0**	**58.3/40.9**	**69.4**	37.9	49.5	30.9
Byte-level T5+LASC$_{Small}$	47M	64.9/53.6	58.7/44.2	52.8/35.9	64.3	36.9	49.2	31.6
CHARFORMER$_{Small}$	48M	69.8/59.4	63.2/48.8	56.8/39.8	68.7	**84.8**	**50.9**	**77.1**
Byte-level T5$_{Base}$	200M	75.6/65.4	68.6/54.3	61.8/44.4	69.4	87.1	57.4	80.9
Byte-level T5+LASC$_{Base}$	205M	70.6/59.7	66.8/52.1	58.8/41.1	67.9	84.8	55.2	79.0
CHARFORMER$_{Base}$	203M	**75.9/65.6**	**70.2/55.9**	**62.6/44.9**	**71.1**	**87.2**	**57.6**	**81.6**
CHARFORMER$_{Tall}$	134M	79.1/68.8	73.6/59.0	66.3/48.5	72.2	88.2	66.6	**85.2**
CHARFORMER$_{Tall,LongPT}$	134M	**81.2/71.3**	**74.2/59.8**	**67.2/49.4**	**72.8**	**88.6**	**67.8**	83.7

Evaluation – Speed, Peak Mem., Parameters

Model	L	d_s	$\|\theta\|$	Speed (steps/s)	FLOPS	Peak Mem.
T5$_{Small}$ (Subword)	512	-	77.1M	28	3.6×10^{12}	-
T5$_{Base}$ (Subword)	512	-	220M	9.3	1.1×10^{13}	-
Byte-level T5$_{Small}$	1024	1	45.0M	29	7.2×10^{12}	989MB
Byte-level T5+LASC$_{Small}$	1024	4	47.4M	55	2.5×10^{12}	530MB
Funnel T5$_{Small}$	1024	*	45.1M	38	4.2×10^{12}	740MB
CHARFORMER$_{Small}$	1024	2	48.5M	32	3.5×10^{12}	765MB
CHARFORMER$_{Small}$	1024	3	48.5M	56	2.6×10^{12}	528MB
CHARFORMER$_{Small}$	1024	5	48.5M	68	1.9×10^{12}	507MB
Byte-level T5$_{Base}$	1024	1	200M	8.2	2.9×10^{13}	3.09GB
Byte-level T5+LASC$_{Base}$	1024	4	205M	15	9.9×10^{12}	1.62GB
CHARFORMER$_{Base}$	1024	2	206M	11	1.6×10^{13}	1.95GB
CHARFORMER$_{Base}$	1024	3	203M	15	1.1×10^{13}	1.63GB
CHARFORMER$_{Tall}$	1024	2	134M	14	1.3×10^{13}	1.73GB
CHARFORMER$_{Tall}$	1024	3	134M	20	8.7×10^{12}	1.34GB

Latent Subwords