Learning Multiscale Transformer Models for Sequence Generation

Bei Li[1] Tong Zheng[1] Yi Jing[1] Chengbo Jiao[2] Tong Xiao[1,2] Jingbo Zhu[1,2]

[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China

International Conference on Machine Learning, 2022

Background

- Transformers have achieved remarkable success on a wide range of tasks in NLP. It can model the relationship between any input tokens. The input consists of a series of words and sub-words.

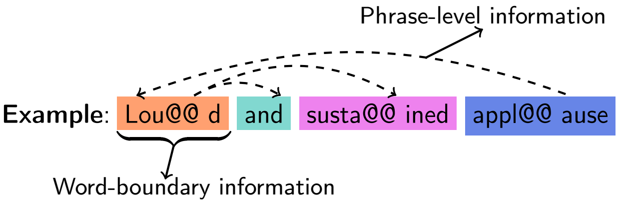
Vanilla Transformer: `Lou@@ d and susta@@ ined appl@@ ause`

Background

- Transformers have achieved remarkable success on a wide range of tasks in NLP. It can model the relationship between any input tokens. The input consists of a series of words and sub-words.

Vanilla Transformer: Lou@@ d and susta@@ ined appl@@ ause

- Despite great potential on most of NLP tasks, the Transformer backbones still have a major shortcoming that it ignores the word-boundary information and other priors, e.g. phrase-level knowledge.



Definition of Scale in NLP

- We redefine the scale from the linguistic perspective in this work (sub-words, words and phrases).
- Sub-words are the lowest-level scale while the phrases are the highest-level scale.

Example: Loud and sustained applause

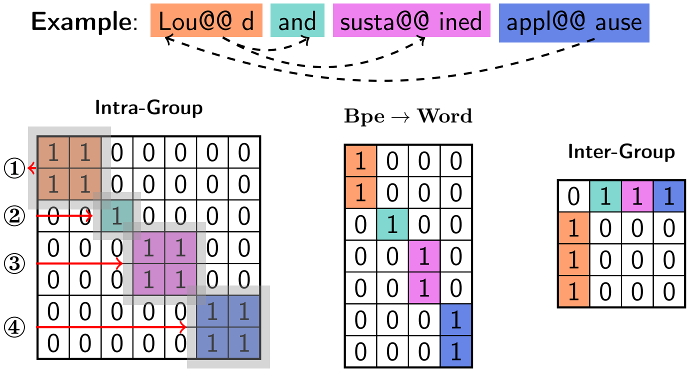
Phrase: Lou@@ d | and | susta@@ ined | appl@@ ause

Word: Lou@@ d | and | susta@@ ined | appl@@ ause

BPE: Lou@@ | d | and | susta@@ | ined | appl@@ | ause

Interactions among Scales

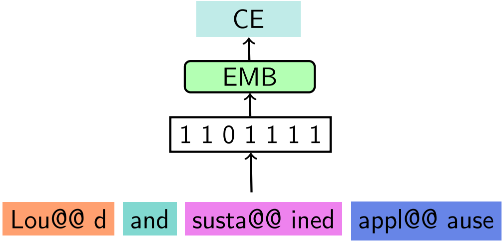
- We establish the relationship among different scales.
  - We regard a sub-word as an individual (Figure (a), each column), and a word as a group ((Figure (a), ①-④)).
  - Intra-group interaction and Inter-group interaction.

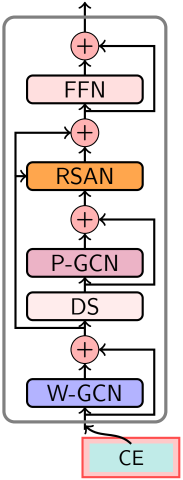


(a) $\mathcal{A}_w$ (b) $\mathcal{G}_{b \to w}$ (c) $\mathcal{A}_p$

UMST

- Class Embedding (CE)





Universal Multiscale Transformer

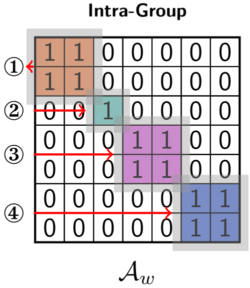
EMB

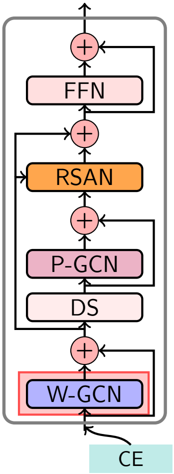
Initialized by a normal distribution, where

$$\sigma = \frac{1}{\sqrt{d}}, \quad \mu = 0 \tag{1}$$

- W-GCN
  - We adopt W-GCN to model the intra-group interaction:

$$\mathrm{GCN}_{\mathrm{word}} = \sigma(\tilde{D}_w^{-\frac{1}{2}} \tilde{\mathcal{A}}_w \tilde{D}_w^{-\frac{1}{2}} \cdot x W_w)$$





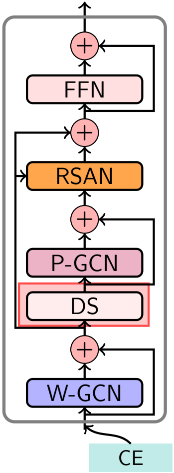
Universal Multiscale Transformer

- P-GCN
  - We adopt a down-sampling operation via $\mathcal{G}_{b\to w}$ to generate word-level representation.

Bpe → Word

1	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1

(a) $\mathcal{G}_{b\to w}$



Universal Multiscale Transformer

- P-GCN
  - We adopt a down-sampling operation via $\mathcal{G}_{b \to w}$ to generate word-level representation.
  - We adopt P-GCN via $\mathcal{A}_p$ to model the inter-group interactions:

$$\mathrm{GCN}_{\mathrm{phrase}} = \sigma(\tilde{D}_p^{-\frac{1}{2}} \tilde{\mathcal{A}}_p \tilde{D}_p^{-\frac{1}{2}} \cdot x W_p)$$

Bpe $\to$ Word

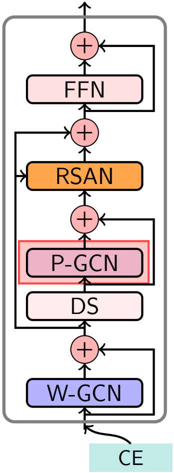
1	0	0	0
1	0	0	0
0	1	0	0
0	0	1	0
0	0	1	0
0	0	0	1
0	0	0	1

(a) $\mathcal{G}_{b \to w}$

Inter-Group

0	1	1	1
1	0	0	0
1	0	0	0
1	0	0	0

(b) $\mathcal{A}_p$

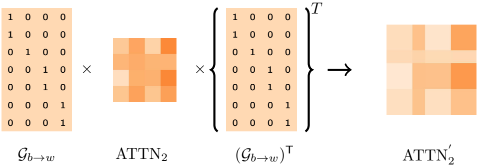


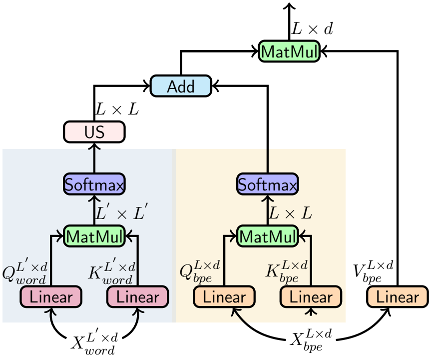
Universal Multiscale Transformer

UMST

- Rectified Self-attention (RSAN)
  - We fuse the multi-scale information in a two-branch manner.
  - To mitigate the gap among different scales:

$$\mathrm{ATTN}_2' = \mathcal{G}_{b\to w} \cdot \mathrm{ATTN}_2 \cdot (\mathcal{G}_{b\to w})^{\mathsf{T}}$$





RSAN module

Benefits

1) Retain information during transformation.
2) Guarantee the normalization.

Results of Machine Translation

- Our UMST outperforms Transformer by 0.88 and 0.44 BLEU points on the base and big configurations, respectively.
- UMST is orthogonal to previous local modeling method e.g. RPR.

Table: Results on the WMT En-De task.

Model	Base		Big	
	Param	**BLEU**	**Param**	**BLEU**
Transformer (Vaswani et al., 2017)	65M	27.30	213M	28.40
Scaling NMT (Ott et al., 2018)	-	-	210M	29.30
DLCL (Wang et al., 2019)	62M	27.30	-	-
MUSE (Zhao et al., 2019)	-	-	-	29.90
MG-SA (Hao et al., 2019)	89M	28.28	272M	29.01
Transformer †	65M	27.63	216M	29.31
MUSE† (Zhao et al., 2019)	68M	27.97	233M	29.11
MSMSA† (Guo et al., 2020)	65M	27.57	233M	28.84
TNT† (Han et al., 2021)	83M	28.48	-	-
UMST	70M	28.51	242M	29.75
UMST + RPR	70M	28.90	242M	30.15

Results of Abstractive Summarization

- Similarly, UMST outperforms the standard Transformer by a large margin.
- The model can still attain nearly 1 rouge gains in terms of three metrics when removing the phrase-level prior knowledge, which demonstrates the essential of word-boundaries.

Table: Results on the CNN-DailyMail dataset.

Model	**RG-1**	**RG-2**	**RG-L**
DynamicConv (Wu et al., 2019)	39.84	16.25	36.73
Bottom-Up (Gehrmann, Deng, and Rush, 2018)	41.22	18.68	38.34
Surface (Liu et al., 2020)	41.00	18.30	37.90
Dman (Fan et al., 2021)	40.98	18.29	37.88
Transformer†	40.55	17.81	37.47
UMST w/o inter-group interactions	41.62	18.65	38.28
UMST	**41.82**	**18.91**	**38.54**

Ablation Study

- Removing any module results in an obvious performance degradation.
- GCN is superior to the GAT and Pooling to model the interactions.

Table: Ablation study on the WMT En-De testset.

Model	**Depth**	**BLEU**	**Depth**	**BLEU**
Transformer	6-6	27.63	12-6	28.67
UMST	6-6	28.51	12-6	29.49
w/o class-embedding	6-6	28.39	12-6	28.99
w/o intra-group interactions	6-6	27.87	12-6	failed
w/o inter-group interactions	6-6	28.06	12-6	29.37
replace GCN with pooling	6-6	27.96	12-6	28.89
replace GCN with GAT	6-6	28.11	12-6	failed

Effect of Encoder Depth and BPE Operations

- UMST beats Transformer under all configurations, attaining almost a 0.76 BLEU gap in average.
- Sentences are likely to be separated into sub-tokens when a vocabulary gets smaller.
- The word boundary information is more essential within a small vocabulary, where UMST can gain more benefits.

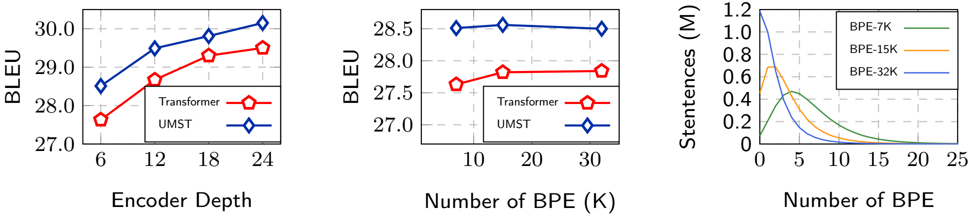
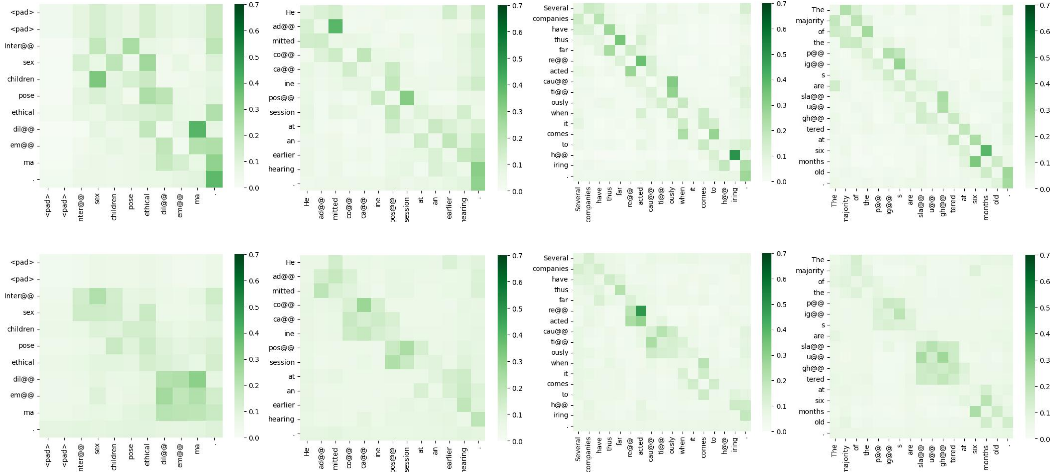


Figure: The comparison of BLEU against different encoder depths and BPE merging operations.

Visualization



Thanks!

Thanks for your attention!

Codebase: https://github.com/libeineu/UMST

Our team: https://github.com/NiuTrans

Any questions please contact with libei_neu@outlook.com

NEU NLP SINCE 1973 小牛翻译 NiuTrans.com