Adaptable Multi-Domain Language Model for Transformer ASR

Taewoo Lee, Min-Joong Lee, Tae Gyoon Kang, Seokyeoung Jung, Minseok Kwon, Yeona Hong, Jungin Lee, Kyoung-Gu Woo, Ho-Gyeong Kim, Jiseung Jeong, Jihyun Lee, Hosik Lee, Young Sang Choi

Samsung Electronics, South Korea

ICASSP 2021

Outline

- Motivation
- Previous Studies
- Proposed Method
- Experimental Results

Motivation

- Catastrophic Forgetting Problem
  - knowledge learned from previous training data disappears from the model when data is sequentially trained in a neural network.

- One simple way to avoid this problem is that to retrain the model from scratch with the newly added data.
  - Drawback : Inefficiency. It takes too long time to pretrain the model.

- Previous studies to relieve the problem
  - **Learning without forgetting (LWF)** adds output logits of previous stage networks to logits of current stage networks [Z. Li 2016].
  - **Elastic weight consolidation (EWC)** constrains weight updates by valuing which weight are important for a task [Kirkpatrick 2017]
  - **Progressive neural networks** avoid forgetting by preserving task specific networks [Rusu 2016]

- However, since these methods require quite a bit of memory or additional procedures, we need a simpler and more effective method.

Previous Study (Adapter in NLP)

- A method to fine-tune BERT models by adding only adjustable 3.6% of parameters [Houlsby 2019]. The method adds small size adapters to the self-attention (SA) and feed forward network (FFN) layers of Transformer, respectively.

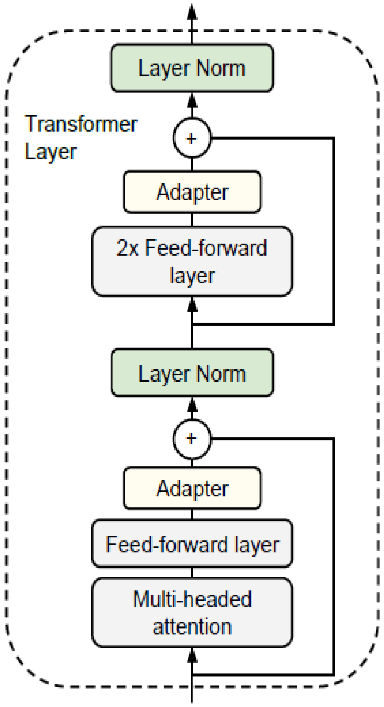


Fig. 1

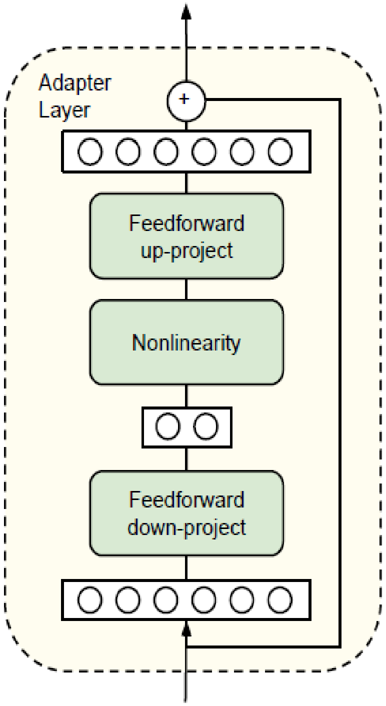


Fig. 2

1. (NLP) Vaswani, Ashish, et al. "Attention Is All You Need."
2. (ASR) L. Dong, S. Xu and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition,“ ICASSP 2018

Previous Study (Adapter in NLP)

- A method to fine-tune BERT models by adding only adjustable 3.6% of parameters [Houlsby 2019]. The method adds small size adapters to the self-attention (SA) and feed forward network (FFN) layers of Transformer, respectively.

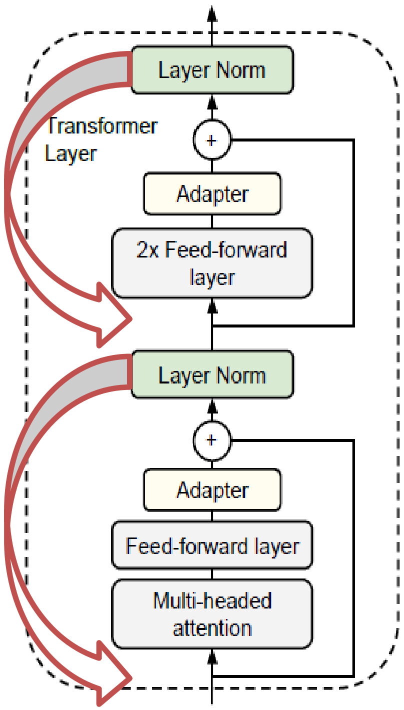


Fig. 1

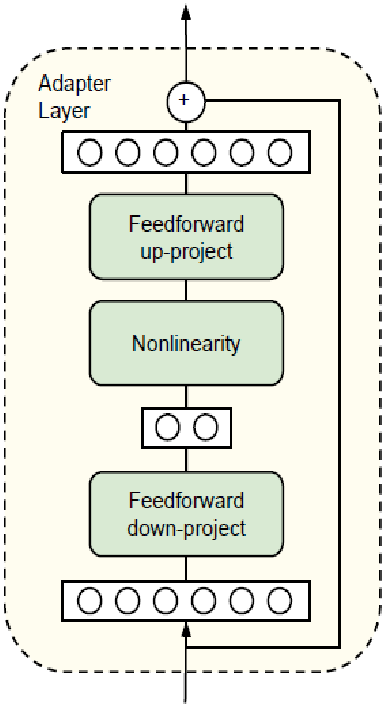


Fig. 2

1. (NLP) Vaswani, Ashish, et al. "Attention Is All You Need."
2. (ASR) L. Dong, S. Xu and B. Xu, "Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition,“ ICASSP 2018

Transformer ASR with External Transformer LMs

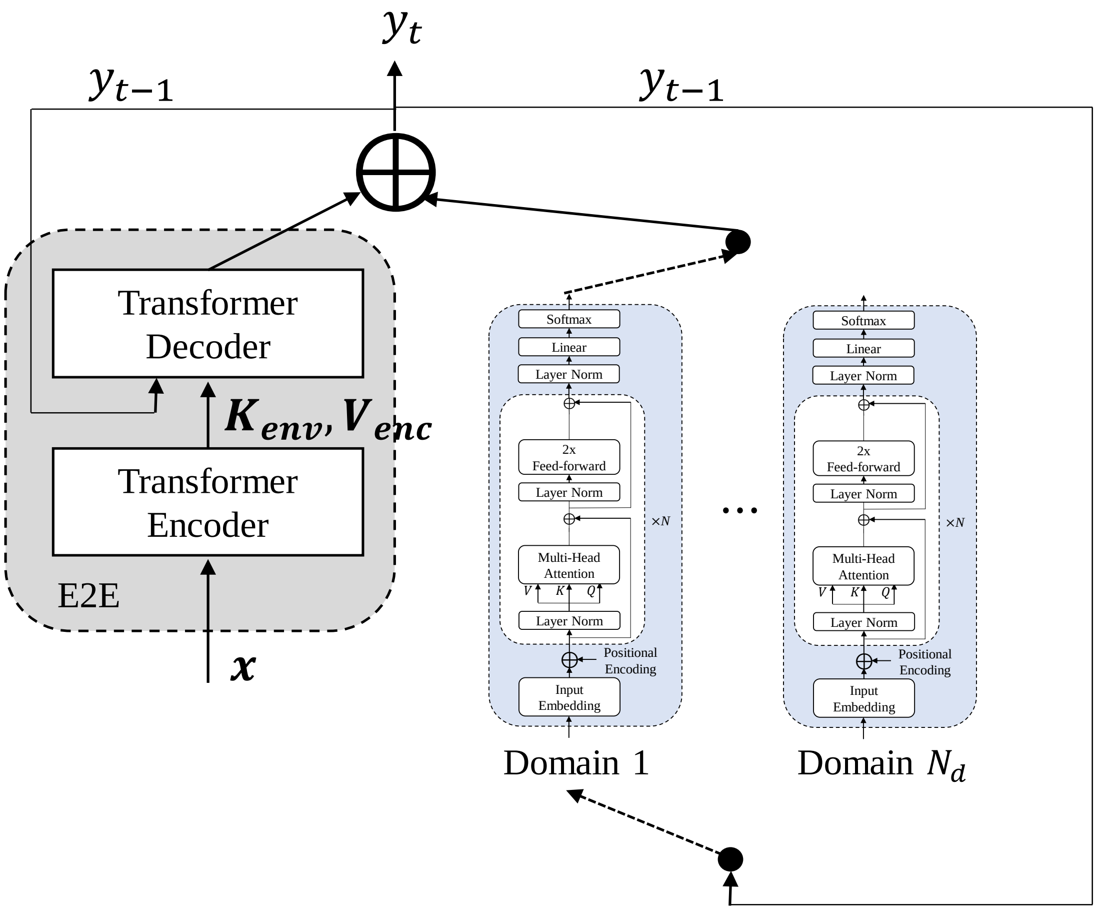


Fig. 3. The dotted line box shows transformer-based E2E ASR model, including encoder and decoder. An external LM is incorporated at each step of beam search

Proposed Method

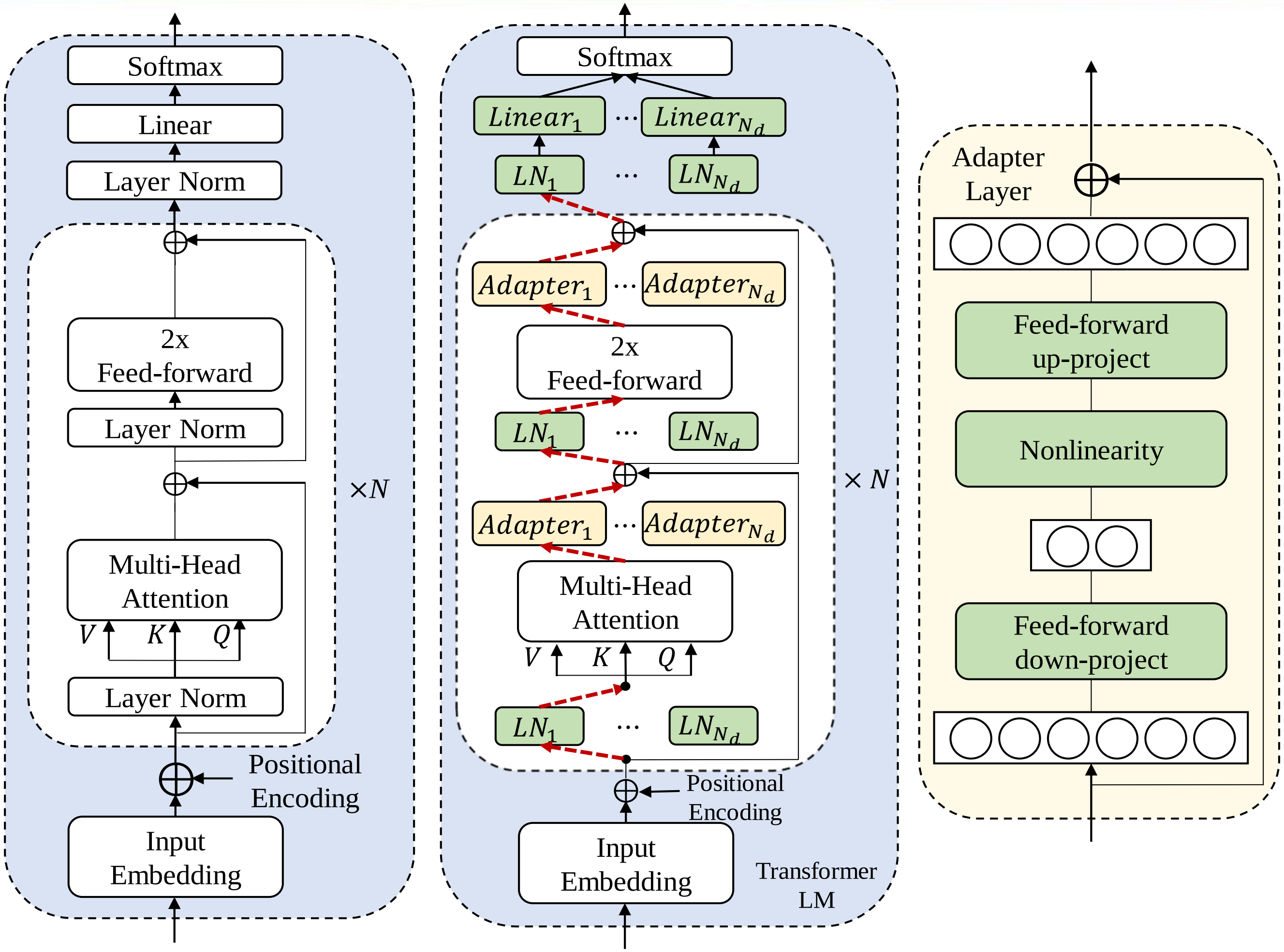
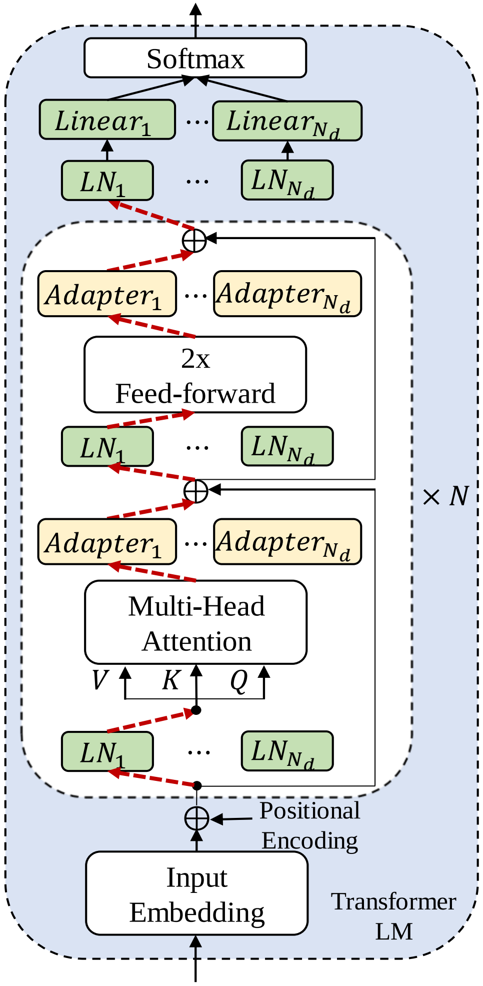


Fig. 4. (**Left**) is an architecture of previous transformer LM. (**Center**) is an architecture of transformer multi-domain LM. In a LM decoder, the adapter module (**Right**) is added on top of multi-head attention and feed-forward layers.

Proposed Method



Advantages

1. Save the number of model parameters.
   - This is because only the parameters required for decoding for each domain can be managed separately through the adapter. This can be useful where memory usage is limited, such as in an on-device environment.

2. Possible to prevent common layers from forgetting previously learned knowledge.
   - This is because learning about newly added data is performed only for the adapter-related layers for each domain.

3. Save time to train model.
   - 3 days (build from scratch) to 1 hour (adapter adaptation)

Experiments (Training data & Test Cases)

1. Training Data
   - The General LM is trained on anonymized 24GB normalized Korean text data consisting of 353M utterances.
   - The Music LM is trained on normalized Korean text data consisting of 45M utterances, in which general and music domain (song title and singer name related commands) corpus are mixed.

2. Test Data

	Domain	# Utterances	Contents	Voice
General LM	In	50K	Bixby use-case scenario such as phone and device control commands, daily conversational question and answering.	Male/ Female
	Out	8K	Domain specific utterances. Especially, domains having its own unique proper nouns such as hospital or doctor’s names.	Male/ Female
Music LM	In	610	Well known song titles and singer names.	Male/ Female
	Out	3709	Newly added song titles and singer names.	Male/ Female

H. Kim, et al., “Knowledge Distillation Using Output Errors for Self Attention End-To-End Models,” in ICASSP, 2019.

Experimental Results

Table 1. WERs of E2E, E2E-G-LM, and E2E-G-LM-A on General Domain TCs

TC	E2E	E2E-G-LM	E2E-G-LM-A
In-Domain	2.42	1.82	**1.69**
Out-Domain	10.62	8.18	**2.84**

Table 2. WERs of E2E, E2E-M-LM, and E2E-M-LM-A on Music Domain TCs

TC	E2E	E2E-M-LM	E2E-M-LM-A
In-Domain	8.2	2.68	**2.46**
Out-Domain	12.66	5.43	**4.13**

Table 3. WERs of iterative adapter fine-tuning with M-LM-A on Music Domain TCs

TC	E2E-M-LM	$M1_{iter1}$	$M1_{iter2}$	$M1_{iter3}$
In-Domain	2.68	2.46	1.97	1.81
Out-Domain	5.43	4.13	3.96	3.87

Experimental Results

Table 4. Iterative fine-tuning performance (WER). The results show a G-LM with iterative fine-tuned adapters can be used as a dedicated music LM.

TC	E2E-M-LM	E2E-M-LM-A	E2E-G-LM	E2E-G-LM $-A_{iter1}$	E2E-G-LM $-A_{iter2}$	E2E-G-LM $-A_{iter3}$	WERR (E2E-G-LM-A_{iter3} – E2E-M-LM)	WERR (E2E-G-LM-A_{iter3} – E2E-M-LM-A)
In-Domain	2.68	2.46	4.65	3.82	2.38	2.19	−0.49	−0.27
	>	<						
Out-Domain	5.43	4.13	11.27	5.75	4.75	4.60	−0.83	0.47

Thank you