Adaptable Multi-Domain Language Model for Transformer ASR

Paper ID: 1298

Taewoo Lee[1], Min-Joong Lee[2], Tae Gyoon Kang[2], Seokyeoung Jung[1], Minseok Kwon[1], Yeona Hong[1], Jungin Lee[1], Kyoung-Gu Woo[1], Ho-Gyeong Kim[2], Jiseung Jeong[2], Jihyun Lee[2], Hosik Lee[2], Young Sang Choi[2]

[1]AI R&D Group, [2]Samsung Advanced Institute of Technology, Samsung Electronics, South Korea

{tw1.lee, minjoong.lee, taeg.kang, jihyun.s.lee}@samsung.com

Abstract

We propose an adapter based multi-domain Transformer based language model (LM) for Transformer ASR. The model consists of a big size common LM and small size adapters. The model can perform multi-domain adaptation with only the small size adapters and its related layers. The proposed model can reuse the full fine-tuned LM which is fine-tuned using all layers of an original model. The proposed LM can be expanded to new domains by adding about 2% of parameters for a first domain and 13% parameters for after second domain. The proposed model is also effective in reducing the model maintenance cost because it is possible to omit the costly and time-consuming common LM pre-training process. Using proposed adapter based approach, we observed that a general LM with adapter can outperform a dedicated music domain LM in terms of word error rate (WER).

1. 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 pre-train the model.

2. Previous Study (Adapter in NLP)

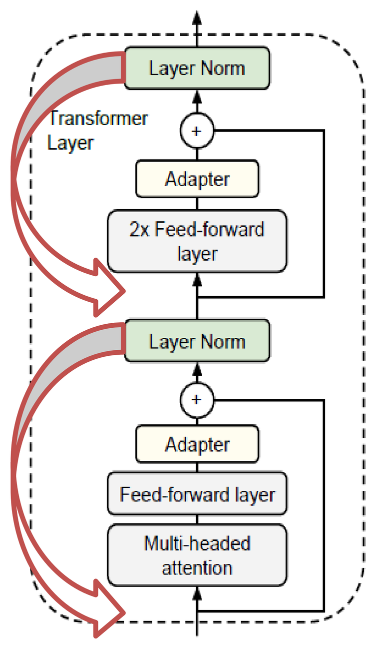

Fig. 1

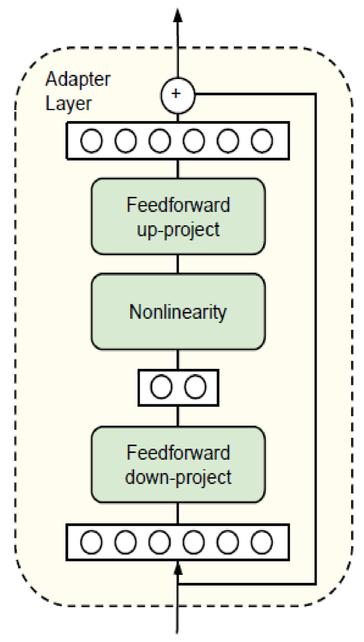

Fig. 2

- Like Resnet block, an adapter module consists of two feedforward layers and one RELU layer like Fig. 2.
- Like Fig 1. The adapter modules are added twice to each Transformer layer. One is added after the projection following multi headed attention and another one is added after two feed forward layers.
- During adapter tuning, the green layers are trained on the downstream data. These layers include the adapter, layer normalization parameters, and final classification layer.
- However, layer structures are slightly different between NLP and ASR Transformers.

3. Proposed Method

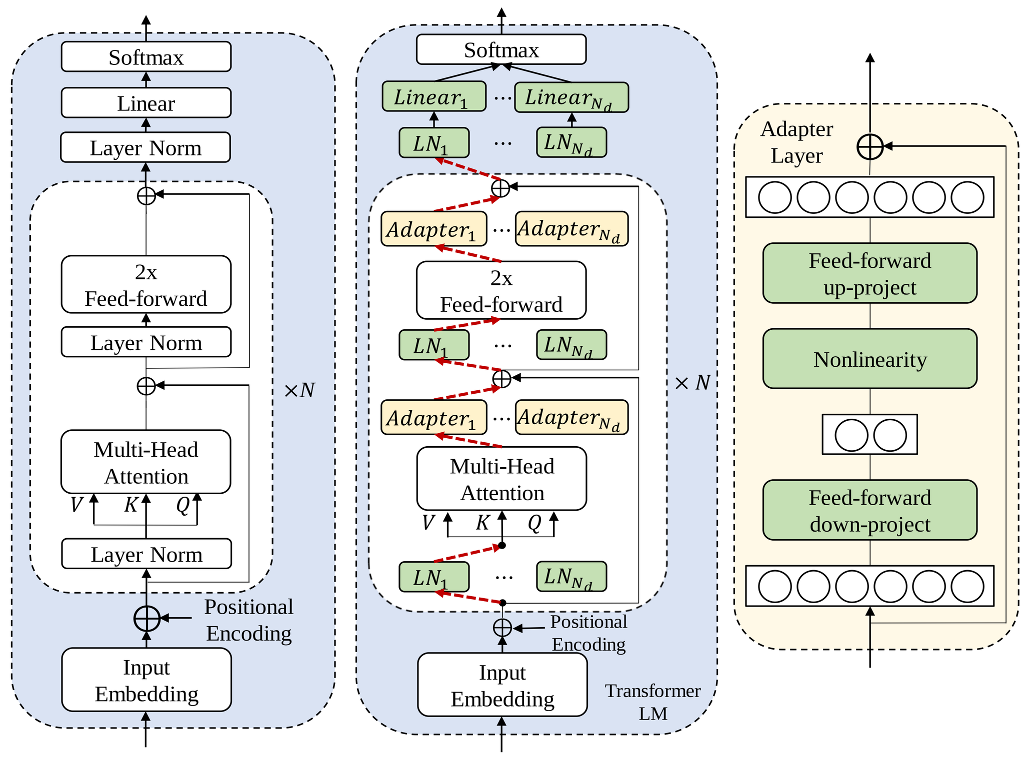

Fig. 3

- Model Architecture
  - The figure on the left is the structure of Transformer LM before the adapter is added.
  - The figure in the middle is the structure of the proposed Transformer multi-domain LM. As shown in the figure, the proposed method is made by adding adapter modules to the existing Transformer LM layers.
  - Finally, the figure on the right is the added adapter module.
- When learning a small amount of new data for each domain, only the adapter, layer normalization, and linear layers indicated in green are trained for the target domain data.
- Even when the domain is expanded, only the adapter-related layers can be branched while leaving the layer corresponding to the large body as it is.

4. Data Set

- Training Data
  - The General LM : anonymized 24GB normalized Korean text data consisting of 353M utterances.
  - The Music LM : normalized Korean text data consisting of 45M utterances
- Test Data

	Domain	# Utterances	Contents
General LM	In	50K	Bixby use-case scenario
	Out	8K	Domain specific utterances. Especially, domains having its own unique proper nouns such as hospital or doctor's names.
Music LM	In	610	Well known song titles and singer names.
	Out	3709	Newly added song titles and singer names.

5. Experiment 1

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

- Table 1 shows how far the recognition rate can be improved when an adapter is applied to a given General LM. By adding an adapter to the LM and further adapting the error sentences obtained from the decoding result, we could get an additional recognition rate improvement over the recognition rate of the already best tuned base model. In particular, out-domain TCs, which included a lot of unique proper nouns, showed a greater improvement.
- Table 2 shows the experimental results for Music LM. As in Table 1, similar results were observed for Music LM.
- Table 3 shows how much the recognition rate can be improved by iterative adapter training. Iterative adapter training refers to a method of repeating the process of training adapter-related layers by extracting error sentences from the decoded result. We have confirmed that the recognition rate is improved up to 3 times, and we were able to further improve the error for a given TC.

6. Experiment 2

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

Table 4.

- Table 4 shows whether the general LM with adapter can also be used as a music LM.
- Music domain TC was used in this experiment.
- Looking at the results in the blue box on the left, you can see that the WERs of the music LM and the music LM with adapter are better than the result of using the general LM. However, if you add an adapter to General LM and repeat iterative adapter training, you can see that you can finally get a lower WER than music LM. Also, you can get lower WER in in-domain than music LM with adapter.

7. Conclusion

1. It can greatly save the number of model parameters.
2. It is possible to prevent common layers from forgetting previously learned knowledge.
3. Since you don't have to train the model from scratch, you can save time to train (adapt) the model.