Supplementary Material for Parameter-Efficient Transfer Learning for NLP

A. Additional Text Classification Tasks

Dataset	Train examples	Validation examples	Test examples	Classes	Avg text length	Reference
20 newsgroups	15076	1885	1885	20	1903	(Lang, 1995)
Crowdflower airline	11712	1464	1464	3	104	crowdflower.com
Crowdflower corporate messaging	2494	312	312	4	121	crowdflower.com
Crowdflower disasters	8688	1086	1086	2	101	crowdflower.com
Crowdflower economic news relevance	6392	799	800	2	1400	crowdflower.com
Crowdflower emotion	32000	4000	4000	13	73	crowdflower.com
Crowdflower global warming	3380	422	423	2	112	crowdflower.com
Crowdflower political audience	4000	500	500	2	205	crowdflower.com
Crowdflower political bias	4000	500	500	2	205	crowdflower.com
Crowdflower political message	4000	500	500	9	205	crowdflower.com
Crowdflower primary emotions	2019	252	253	18	87	crowdflower.com
Crowdflower progressive opinion	927	116	116	3	102	crowdflower.com
Crowdflower progressive stance	927	116	116	4	102	crowdflower.com
Crowdflower US economic performance	3961	495	496	2	305	crowdflower.com
Customer complaint database	146667	18333	18334	157	1046	catalog.data.gov
News aggregator dataset	338349	42294	42294	4	57	(Lichman, 2013)
SMS spam collection	4459	557	558	2	81	(Almeida et al., 2011)

Table 1. Statistics and references for the additional text classification tasks.

Dataset	Epochs (Fine-tune)	Epochs (Adapters)
20 newsgroups	50	50
Crowdflower airline	50	20
Crowdflower corporate messaging	100	50
Crowdflower disasters	50	50
Crowdflower economic news relevance	20	20
Crowdflower emotion	20	20
Crowdflower global warming	100	50
Crowdflower political audience	50	20
Crowdflower political bias	50	50
Crowdflower political message	50	50
Crowdflower primary emotions	100	100
Crowdflower progressive opinion	100	100
Crowdflower progressive stance	100	100
Crowdflower US economic performance	100	20
Customer complaint database	20	20
News aggregator dataset	20	20
SMS spam collection	50	20

Table 2. Number of training epochs selected for the additional classification tasks.

Parameter	Search Space
1) Input embedding modules	Refer to Table 4
2) Fine-tune input embedding module	{True, False}
3) Lowercase text	{True, False}
4) Remove non alphanumeric text	{True, False}
5) Use convolution	{True, False}
6) Convolution activation	{relu, relu6, leaky relu, swish, sigmoid, tanh}
7) Convolution batch norm	{True, False}
8) Convolution max ngram length	{2, 3}
9) Convolution dropout rate	[0.0, 0.4]
10) Convolution number of filters	[50, 200]
11) Convolution embedding dropout rate	[0.0, 0.4]
12) Number of hidden layers	{0, 1, 2, 3, 5}
13) Hidden layers size	{64, 128, 256}
14) Hidden layers activation	{relu, relu6, leaky relu, swish, sigmoid, tanh}
15) Hidden layers normalization	{none, batch norm, layer norm}
16) Hidden layers dropout rate	{0.0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5}
17) Deep tower learning rate	{0.001, 0.005, 0.01, 0.05, 0.1, 0.5}
18) Deep tower regularization weight	{0.0, 0.0001, 0.001, 0.01}
19) Wide tower learning rate	{0.001, 0.005, 0.01, 0.05, 0.1, 0.5}
20) Wide tower regularization weight	{0.0, 0.0001, 0.001, 0.01}
21) Number of training samples	{1e5, 2e5, 5e5, 1e6, 2e6}

Table 3. The search space of baseline models for the additional text classification tasks.

ID	Dataset size (tokens)	Embed dim.	Vocab. size	Training algorithm	TensorFlow Hub Handles Prefix: `https://tfhub.dev/google/`
English-small	7B	50	982k	Lang. model	`nnlm-en-dim50-with-normalization/1`
English-big	200B	128	999k	Lang. model	`nnlm-en-dim128-with-normalization/1`
English-wiki-small	4B	250	1M	Skipgram	`Wiki-words-250-with-normalization/1`
English-wiki-big	4B	500	1M	Skipgram	`Wiki-words-500-with-normalization/1`
Universal-sentence-encoder	-	512	-	(Cer et al., 2018)	`universal-sentence-encoder/2`

Table 4. Options for text input embedding modules. These are pre-trained text embedding tables. We provide the handle for the modules that are publicly distributed via the TensorFlow Hub service (`https://www.tensorflow.org/hub`).

B. Learning Rate Robustness

We test the robustness of adapters and fine-tuning to the learning rate. We ran experiments with learning rates in the range $[2 \cdot 10^{-5}, 10^{-3}]$, and selected the best hyperparameters for each method at each learning rate. Figure 1 shows the results.

Dataset	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
20 newsgroups	Universal-sentence-encoder	False	True	True	False	relu6	False	2	0.37	94	0.38	1	128	leaky relu	batch norm	0.5	0.5	0	0.05	0.0001	1000000
Crowdflower airline	English-big	False	False	False	True	leaky relu	False	3	0.36	200	0.07	0	128	tanh	layer norm	0.4	0.1	0.001	0.05	0.001	200000
Crowdflower corporate messaging	English-big	False	False	True	True	tanh	True	3	0.40	56	0.40	1	64	tanh	batch norm	0.5	0.5	0.001	0.01	0	200000
Crowdflower disasters	Universal-sentence-encoder	True	True	False	True	swish	True	3	0.27	52	0.22	0	64	relu	none	0.2	0.005	0.0001	0.005	0.01	500000
Crowdflower economic news relevance	Universal-sentence-encoder	True	True	False	False	leaky relu	False	2	0.27	63	0.04	3	128	swish	layer norm	0.2	0.01	0.01	0.001	0	100000
Crowdflower emotion	Universal-sentence-encoder	False	True	False	False	relu6	False	3	0.35	132	0.34	1	64	tanh	none	0.05	0.05	0	0.05	0	200000
Crowdflower global warming	Universal-sentence-encoder	False	True	True	False	swish	False	3	0.39	200	0.36	1	128	leaky relu	batch norm	0.4	0.05	0	0.001	0.001	1000000
Crowdflower political audience	English-small	True	False	True	True	relu	False	3	0.11	98	0.07	0	64	relu	none	0.5	0.05	0.001	0.001	0	100000
Crowdflower political bias	English-big	False	True	True	False	swish	False	3	0.12	81	0.30	0	64	relu6	none	0	0.01	0	0.005	0.01	200000
Crowdflower political message	Universal-sentence-encoder	False	False	True	False	swish	True	2	0.36	57	0.35	0	64	tanh	none	0.5	0.01	0.001	0.005	0	200000
Crowdflower primary emotions	English-big	False	True	True	True	swish	False	3	0.40	191	0.03	0	256	relu6	none	0.5	0.1	0.001	0.05	0	200000
Crowdflower progressive opinion	English-big	True	False	True	True	relu6	False	3	0.40	199	0.28	0	128	relu	batch norm	0.3	0.1	0.01	0.005	0.001	200000
Crowdflower progressive stance	Universal-sentence-encoder	True	False	True	False	relu	True	3	0.01	195	0.00	2	256	tanh	layer norm	0.4	0.005	0	0.005	0.0001	500000
Crowdflower us economic performance	English-big	True	False	True	True	tanh	True	2	0.31	53	0.24	1	256	leaky relu	batch norm	0.3	0.05	0.0001	0.001	0.0001	100000
Customer complaint database	English-big	True	False	False	False	tanh	False	2	0.03	69	0.10	1	256	leaky relu	layer norm	0.1	0.05	0.0001	0.05	0.001	1000000
News aggregator dataset	Universal-sentence-encoder	False	True	True	False	sigmoid	True	2	0.00	156	0.29	3	256	relu	batch norm	0.05	0.05	0	0.5	0.0001	1000000
Sms spam collection	English-wiki-small	True	True	True	True	leaky relu	False	3	0.20	54	0.00	1	128	leaky relu	batch norm	0	0.1	0	0.05	0.01	1000000

Table 5. Search space parameters (see Table 3) for the AutoML baseline models that were selected.

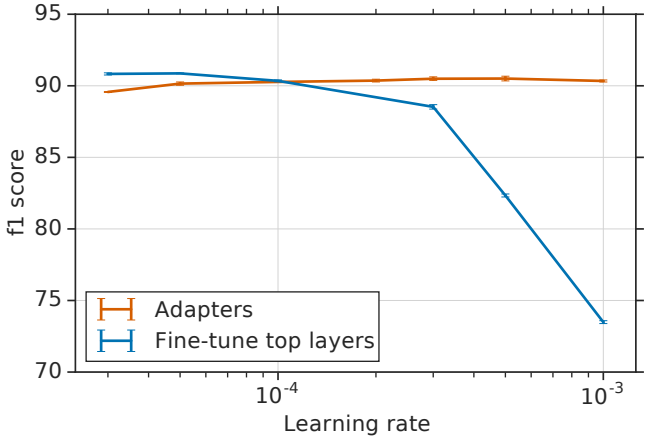

Figure 1. Best performing models at different learning rates. Error vars indicate the s.e.m. across three random seeds.

References

Almeida, T. A., Hidalgo, J. M. G., and Yamakami, A. Contributions to the Study of SMS Spam Filtering: New Collection and Results. In *Proceedings of the 11th ACM Symposium on Document Engineering*. ACM, 2011.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018.

Lang, K. Newsweeder: Learning to filter netnews. In *ICML*, 1995.

Lichman, M. UCI machine learning repository, 2013.