Appendix

A Hyper-parameters

The MultiFit architecture has 4 QRNN layers with a hidden dimensionality of 1550, a vocabulary size of 15,000 subword tokens, and an embeding size of 400. The vocabularies were computed using the SentencePiece[1] unigram language model (Kudo, 2018) with 99% character coverage for Chinese and Japanese and 100% for the rest. The encoder's and decoder's weights are shared (Press and Wolf, 2017). The output of the last QRNN layer (the last time step concatenated with an average and maximum pooled over time steps) is passed to the classifier with 2 dense layers.

Our language models were trained for 10 epochs on 100 million tokens of Wikipedia articles and then fine-tuned for 20 epochs on the corresponding dataset (MLDoc or CLS). The classifier was fine-tuned for 4 to 8 epochs. Results of the best model based on accuracy on the validation set are reported. We used a modified version of 1cycle learning rate schedule (Smith, 2018) that uses cosine instead of linear annealing, cyclical momentum and discriminative finetuning (Howard and Ruder, 2018). Our batch size for language model training was 50 and for classification tasks 18. We were using BPTT of length 70. Due to the large amount of available training data our pretrained language models were trained without any dropout. We used the same dropout values as (Howard and Ruder, 2018) multiplied by 0.3 and 0.5 for fine-tuning of language models and the classification task respectively. We used weight decay of 0.01 for both tasks. The final regularization method was label smoothing (Szegedy et al., 2016) with epsilon of 0.1.

B Speed comparison hyper-parameters

For the speed comparison, we use the same architecture and only change the underlying RNN cell (QRNN or LSTM). We pretrain and fine-tune both models on 15k tokens on a Tesla V100. For pretraining, we use a BPTT size of 70 and a batch size of 64. For fine-tuning, we use a batch size of 32.

[1] https://github.com/google/sentencepiece

References

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of ACL 2018*.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, volume 2, pages 157–163.

Leslie N Smith. 2018. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of CVPR 2016*.