A Detailed training setup

We replicate the model configuration (embed=512, ffn=1024, head=4) as the baseline in (Wu et al., 2019). In addition, the batch size of 8192, attention dropout of 0.1, and relu dropout of 0.1 is used as suggested by Wang et al. (2019). When training the standard Transformer from scratch, we follow the *inverse_sqrt* learning rate schedule with learning rate of 0.0015 and warmup of 8k. To speed up convergence of FDMs (e.g. `MT`, `LayerDrop`), all of them are finetuned from the pre-trained baseline with learning rate of 0.0005 and warmup of 4k.

At inference, we use a beam size of 5 and average last 5 checkpoints. We use case insensitive BLEU score evaluated by *multi-bleu.perl* as previous works.

References

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning deep transformer models for machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy.