

000 A Detailed training setup

001 We replicate the model configuration (embed=512,
 002 ffn=1024, head=4) as the baseline in (Wu et al.,
 003 2019). In addition, the batch size of 8192, attention
 004 dropout of 0.1, and relu dropout of 0.1 is used as
 005 suggested by Wang et al. (2019). When training the
 006 standard Transformer from scratch, we follow the
 007 *inverse_sqrt* learning rate schedule with learning
 008 rate of 0.0015 and warmup of 8k. To speed up
 009 convergence of FDMs (e.g. MT, LayerDrop), all
 010 of them are finetuned from the pre-trained baseline
 011 with learning rate of 0.0005 and warmup of 4k.

012 At inference, we use a beam size of 5 and av-
 013 erage last 5 checkpoints. We use case insensitive
 014 BLEU score evaluated by *multi-bleu.perl* as previ-
 015 ous works.

017 References

019 Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu,
 020 Changliang Li, Derek F. Wong, and Lidia S. Chao.
 021 2019. Learning deep transformer models for ma-
 022 chine translation. In *Proceedings of the 57th Annual*
 023 *Meeting of the Association for Computational Lin-*
guistics, pages 1810–1822, Florence, Italy.

024
 025
 026
 027
 028
 029
 030
 031
 032
 033
 034
 035
 036
 037
 038
 039
 040
 041
 042
 043
 044
 045
 046
 047
 048
 049

050
 051
 052
 053
 054
 055
 056
 057
 058
 059
 060
 061
 062
 063
 064
 065
 066
 067
 068
 069
 070
 071
 072
 073
 074
 075
 076
 077
 078
 079
 080
 081
 082
 083
 084
 085
 086
 087
 088
 089
 090
 091
 092
 093
 094
 095
 096
 097
 098
 099