Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A] No theoretical results.
   (b) Did you include complete proofs of all theoretical results? [N/A] No theoretical results.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Training scripts and model code are included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4, Appendex A.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4
   (b) Did you mention the license of the assets? [Yes] See Section 4
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] No new assets introduced in the work.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] No human subjects.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] No human subjects.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] No human subjects.

A Appendix

A.1 Scheduling ADMM Iterations

ADMM can converge effectively with as few as 80 training steps in each ADMM iteration. For example, RTE (2.5k training samples), ADMM successfully converges with a minibatch of 32 and one ADMM iteration per epoch. However, increasing the number of training steps between iterations can reduce reliance on a high learning rate. Note that a high learning rate is necessary to allow the optimizer to relatively quickly push larger parameters towards 0 in a reasonable number of training steps, since the practical parameter delta in a single training step is proportional to the product of the learning rate and ρ. Furthermore, a small learning rate reduces the effectiveness of the regularizer and decreases model similarity.

Experimentally, ADMM will achieve its maximum accuracy once 10 ADMM iterations have occurred. However, further optimizing, does not appear to harm model accuracy. While further training is typically not desirable for small tasks — training is frequently extended for these tasks to have a sufficiently large training period each ADMM iteration — for large tasks tens of ADMM iterations may be performed such that the fine-tune can continue for sufficient time. For example, a fine-tune on QNLI for just 3 epochs may perform nearly 50 ADMM iterations.

Table 3: NxMTransformer Training Hyperparameters. Smaller tasks utilize larger learning rates and penalty parameters (ρ) since ADMM iterations for these tasks are much shorter (See Appendix A.1).

Tasks	Learning Rates	ρ	Batch Size	Epochs
MNLI, QNLI, SST-2	1e-5, 3e-5, 5e-5	4e-4, 1e-3, 3e-3	16, 32	5
CoLA, STS-B, MRPC, RTE	5e-5, 7e-5, 9e-5, 1e-4	3e-3, 6e-3, 1e-2	16, 32	10