A Training Details

Table 1: Hyperparameters for BERT and RoBERTa in all downstream tasks. LR: learning rate; BSZ: batch size; EP: training epochs; WP: warmup proportion; MSL: maximum sequence length; GLR: generator learning rate; DS: dropout step.

	LR	BSZ	EP	WP	MSL	GLR	DS
SST-2	1e-5	32	3	0.06	128	5e-5	12
MRPC	2e-5	32	10	0.06	128	5e-5	12
QNLI	1e-5	32	3	0.06	128	5e-5	12
MNLI-mm	3e-5	32	3	0.1	128	5e-5	15
CoLA	1e-5	16	10	0.06	128	5e-5	10
IMDB	3e-5	32	10	0.1	256	5e-5	12
CoNLL03	5e-5	32	5	0.1	128	1e-4	8
PTB	5e-5	32	5	0.1	128	1e-4	8
SWAG	2e-5	16	3	0	128	5e-5	15

B Resource Usage

Table 2: Resource Usage for AttendOut on RoBERTa-base. MS: model size; GS: generator size; MU: memory usage; ET: epoch time.

	MS	GS	MU	ET
RoBERTa	119M	-	10.6G	8.5mins
RoBERTa + AttendOut	119M	40M	18.9G	18mins