On the Effectiveness of Intermediate Training on an Uncurated Collection of Datasets.
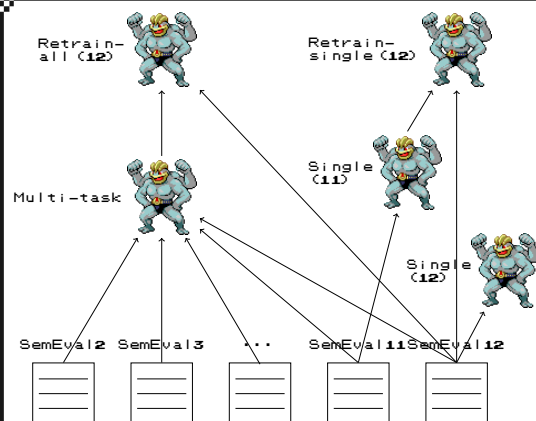
Data

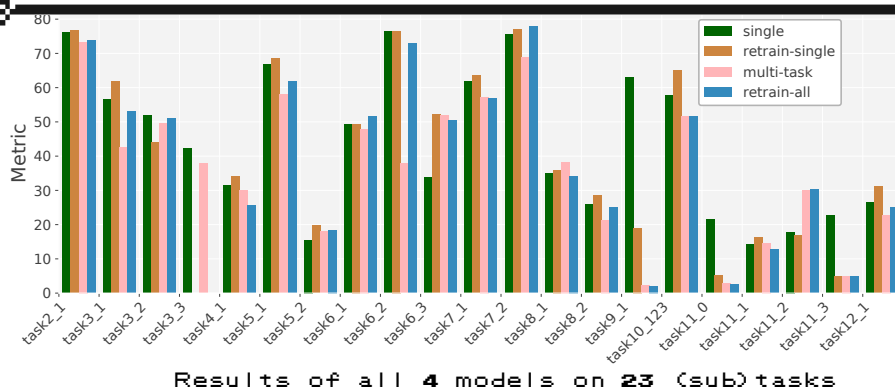| Name	Subtasks	Languages	Size
2. MultiCoNER II	NER	BN, DE, EN, ES, FA, FR, HI, IT, PT, SV, UK, ZH	2,672,490
3. News persuasion	1. News categorization	EN, FR, GE, IT, PO, RU	741,561
	2. Framing classification	EN, FR, GE, IT, PO, RU	725,740
	3. Persuasion technique classification	EN, FR, GE, IT, PO, RU	19,561,550
4. ValueEval	Human value classification	EN	116,294
5. Clickbait spoiling	1. Spoiler type classification	EN	34,520
	2. Spoiler detection	EN	1,647,176
6. LegalEval	1. Rhetorical role detection	EN	755,280
	2. NER	EN	369,205
	3. Legal judgement prediction	EN	5,082
7. Clinical NLI	1. Entailment	EN	21,828
	2. Evidence retrieval	EN	311,687
8. Medical claims	1. Claim identification	EN	549,231
	2. PIO frame extraction	EN	78,864
9. Tweet intimacy	Intimacy Analysis	EN, ES, IT, PO, FR, ZH	73,698
10. Explainable sexism	1. Sexism detection	EN	262,939
	2. Sexism classification	EN	68,043
	3. Fine-grained sexism classification	EN	68,043
11. Le-Wi-Di	1. Hate speech detection*	EN	14,252
	2. Misogyny detection*	AR	12,788
	3. Abuse detection*	EN	64,738
	4. Offensiveness detection*	EN	145,245
12. AfriSenti-SemEval	Sentiment classification	AM, DZ, HA, IG, KR, MA, PCM, PT, SW, TS, TWI, YO	795,449

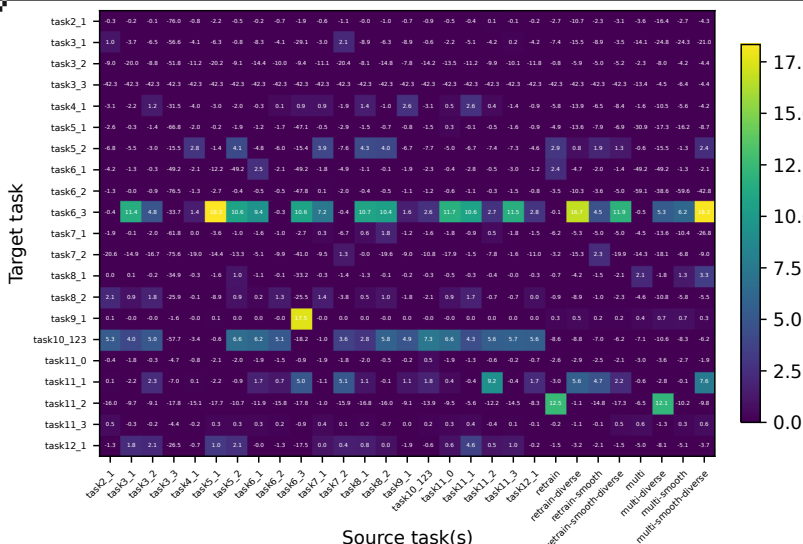Overview of all text-based SemEval 2023 tasks

Models

Overview of single-task baselines and multi-task models for task 12

Results

Results of all 4 models on 23 (sub)tasks

Transfer

Performance difference to baseline when transferring from 1/all tasks to a single target task

Correlations

Feature	Pearson
Baseline src task	0.009
Baseline tgt task	-0.109
src base/tgt base	0.093
Overlap tasktypes	-0.116
num src langs	-0.079
num tgt langs	-0.111
lang overlap	0.022
dom overlap	0.088
src size	-0.500
tgt size	-0.562
src size/tgt size	-0.133

Correlations of performance difference to baseline with specific features.

Takeaways

► Non-trivial to obtain improvements with multi-task/intermediate training
► Target task more consistent than source
► Highest correlations for dataset size
► No substantial gains from heterogeneous batches or scaling up