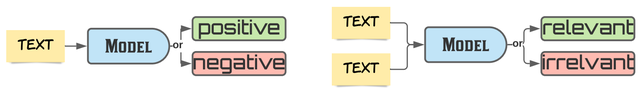
# Does QA-based intermediate training help fine-tuning LM for text classification?

Shiwei Zhang, Xiuzhen (Jenny) Zhang
School of Computing Technologies, RMIT University, Australia

## Text Classification

**Text Classification:** is the problem of classifying text into categories or classes.




## Intermediate Training for Text Classification

- (Pruksachatkun et al., 2020): NLI and QA are the best.
- (Vu et al., 2020): SQuAD-2.0 is the most favourable.

**However, only a few text classifications tasks and one language model involved in their experiments.**

## Using SQuAD-2.0 as the intermediate task for Text Classification

- First, load a pre-trained LM and add a span classification head on top of it, and train it on SQuAD-2.0.
- Next, switch to a sequence classification head, and fine-tune it on downstream text classification tasks.

## Performance for single-sequence text classificat

| | AGNEWS | SST2 | LIAR | OFFENSIVE | HATE | COLA | EMOTION | IRONY |
|---|---|---|---|---|---|---|---|---|
| ELECTRA | 94.46 | 94.61 | 26.63 | 83.48 | 48.01 | 67.65 | 82.59 | 71.96 |
| ELECTRA(T) | 94.59$^{+}$ | 94.26$^{-}$ | 27.76$^{+}$ | 82.91$^{-}$ | 44.90$^{-}$ | 67.01$^{-}$ | 81.86$^{-}$ | 70.96$^{-}$ |
| RoBERTa | 94.84 | 93.00 | 27.65 | 83.18 | 44.19 | 58.84 | 82.75 | 71.41 |
| RoBERTa(T) | 94.82$^{=}$ | 94.15$^{+}$ | 27.35$^{-}$ | 83.45$^{+}$ | 46.62$^{+}$ | 57.17$^{-}$ | 81.79$^{-}$ | 69.35$^{-}$ |
| MobileBERT | 94.57 | 90.13 | 26.07 | 84.71 | 43.66 | 49.99 | 78.23 | 63.08 |
| MobileBERT(T) | 94.32$^{-}$ | 91.05$^{+}$ | 26.27$^{+}$ | 85.01$^{+}$ | 45.57$^{+}$ | 50.25$^{+}$ | 79.72$^{+}$ | 62.36$^{-}$ |
| SqueezeBERT | 94.68 | 89.90 | 27.26 | 84.09 | 41.97 | 44.50 | 78.72 | 66.07 |
| SqueezeBERT(T) | 94.09$^{-}$ | 89.10$^{-}$ | 27.72$^{+}$ | 83.61$^{-}$ | 40.54$^{-}$ | 35.37$^{-}$ | 77.73$^{-}$ | 66.44$^{+}$ |

Table 2: Performance(%) for single-sequence text classification tasks. Models with SQuAD2.0 intermediate tuning are denoted with T, $+$, $=$ and $-$ denote increase, equal and decrease in performance for SQuAD-tuned models.

***SQuAD2-tuned models for single-sequence text classification tasks have mixed results.***

## Performance for pairwise classification tasks.

| | QQP | QNLI | WNLI | MNLI | WIKIQA | BOOLQ | MRPC | RTE |
|---|---|---|---|---|---|---|---|---|
| ELECTRA | 91.69 | 92.09 | 47.88 | 88.52 | 46.04 | 84.16 | 88.60 | 77.61 |
| ELECTRA(T) | 91.45$^{-}$ | **92.44**$^{+}$ | 52.58$^{+}$ | 88.77$^{+}$ | **50.43**$^{+}$ | **86.34**$^{+}$ | 87.78$^{-}$ | 78.34$^{+}$ |
| RoBERTa | 91.24 | 92.04 | 56.34 | 87.69 | 43.41 | 84.22 | 89.56 | 75.33 |
| RoBERTa(T) | 91.14$^{-}$ | **92.42**$^{+}$ | 56.34$^{=}$ | 87.65$^{=}$ | **52.45**$^{+}$ | **84.54**$^{+}$ | 88.31$^{-}$ | 79.18$^{+}$ |
| MobileBERT | 89.09 | 89.18 | 46.48 | 82.63 | 40.18 | 77.65 | 83.69 | 56.68 |
| MobileBERT(T) | 88.94$^{-}$ | **90.88**$^{+}$ | 35.21$^{-}$ | 82.45$^{-}$ | **52.60**$^{+}$ | **81.63**$^{+}$ | 86.87$^{+}$ | 67.75$^{+}$ |
| SqueezeBERT | 89.32 | 89.16 | 52.11 | 80.49 | 41.70 | 79.45 | 83.62 | 68.11 |
| SqueezeBERT(T) | 89.07$^{-}$ | **90.13**$^{+}$ | 39.90$^{-}$ | 80.05$^{-}$ | **50.89**$^{+}$ | **79.98**$^{+}$ | 85.31$^{+}$ | 66.79$^{-}$ |

Table 3: Performance(%) for pairwise classification tasks. Models with SQuAD2.0 intermediate tuning are denoted with T, where $+$, $=$ and $-$ denote increase, equal and decrease in performance for SQuAD-tuned models. Note the positive transfer results on QA tasks QNLI, WIKIQA and BOOLQ.

***SQuAD2-tuned models have consistently better performance for QA tasks.***

## Conclusion

- SQuAD2-tuned models do **NOT** have better performance, whether single-sequence or sequence-pair, or data-rich or data-poor settings, which suggests that high-level inference intermediate tasks may not generally produce positive transfer as previously thought.
- SQuAD2-tuned models are more likely to have positive transfer results for **QA tasks**, which suggests further research is needed to investigate if task similarity rather than task complexity plays a significant role for intermediate training.

## Contact

xiuzhen.zhang@rmit.edu.au, dr.shiwei.zhang@gmail.com

## Acknowledgements

This initiative was funded by the Australian government Department of Defence and the Office of National Intelligence under the AI for Decision Making Program, delivered in partnership with the Defence Science Institute in Victoria.