Paying Per-label Attention for Multi-label Extraction from Radiology Reports (Supplementary Material)

Patrick Schrempf[1,2], Hannah Watson[1], Shadia Mikhael[1], Maciej Pajak[1], Matúš Falis[1], Aneta Lisowska[1], Keith W. Muir[3], David Harris-Birtill[2], and Alison Q. O'Neil[1,4]

[1] Canon Medical Research Europe, Edinburgh, United Kingdom
[2] University of St Andrews, United Kingdom
[3] Institute of Neuroscience & Psychology, University of Glasgow, United Kingdom
[4] University of Edinburgh, United Kingdom
patrick.schrempf@eu.medical.canon

Model	**#Parameters**	**Training time [s]**	**Inference time [s/sample]**
BoW + RF	n/a	14_{1}	$0.2933_{0.0040}$
Word2Vec	166,524	46_{8}	$0.0022_{0.0001}$
CAML [2]	1,021,176	250_{43}	$0.0099_{0.0008}$
Bi-GRU	2,889,852	111_{32}	$0.0066_{0.0003}$
Bi-GRU + single attention	3,371,132	120_{55}	$0.0062_{0.0003}$
Bi-GRU + per-label attention	3,401,852	376_{127}	$0.0109_{0.0004}$
BERT	109,577,596	1115_{322}	$0.0565_{0.0025}$
BioBERT	109,577,596	927_{171}	$0.0575_{0.0008}$
ALARM + softmax	109,458,556	911_{243}	$0.0590_{0.0013}$
ALARM + per-label attention	125,233,276	1448_{375}	$0.0740_{0.0002}$

Table 1: Number of parameters, training time (over 838 samples) and inference time (per sample) for all models. All timings are given as $\text{mean}_{\text{standard deviation}}$ of 5 runs with different random seeds. The fastest model to train is the random forest model. The Bi-GRU network is significantly faster to train than BERT [1] and ALARM [3] due to the smaller number of parameters. The only model that is faster than the Bi-GRU model is Word2Vec which has a far inferior F1 score. The random forest model is the slowest at inference time because it has n_L models (one model per label) - the inference could be parallelised to improve performance.

Model	Embedding	Data	All	Negative	Uncertain	Positive
Bi-GRU	MIMIC	S	$0.584_{0.022}$	$0.496_{0.089}$	$0.204_{0.031}$	$0.642_{0.012}$
Bi-GRU	MIMIC	N-S	$0.908_{0.004}$	$0.956_{0.004}$	$0.427_{0.058}$	$0.927_{0.004}$
Bi-GRU	Random	N+S	$0.893_{0.002}$	$0.962_{0.008}$	$0.432_{0.033}$	$0.903_{0.002}$
Bi-GRU	MIMIC	N+S	$0.921_{0.003}$	$\mathbf{0.970}_{0.006}$	$0.573_{0.011}$	$0.932_{0.004}$
ALARM	MIMIC	S	$0.569_{0.028}$	$0.725_{0.062}$	$0.128_{0.041}$	$0.531_{0.028}$
ALARM	MIMIC	N-S	$0.906_{0.011}$	$0.944_{0.005}$	$0.532_{0.087}$	$0.923_{0.010}$
ALARM	MIMIC	N+S	$\mathbf{0.928}_{0.008}$	$0.965_{0.004}$	$\mathbf{0.689}_{0.039}$	$\mathbf{0.936}_{0.008}$

Table 2: Results for our ablation studies showing *micro-averaged* F1 as $\text{mean}_{\text{standard deviation}}$ of 5 runs with different random seeds (all models are trained with per-label attention). N data is the NHS GGC dataset and S is the synthetic dataset. "All" combines the classes "negative", "uncertain" and "positive". Bold indicates the best model for each metric.

Model	Embedding	Data	All	Negative	Uncertain	Positive
Bi-GRU	MIMIC	S	$0.400_{0.039}$	$0.504_{0.050}$	$0.106_{0.029}$	$0.590_{0.065}$
Bi-GRU	MIMIC	N-S	$0.551_{0.024}$	$0.623_{0.024}$	$0.268_{0.089}$	$0.761_{0.026}$
Bi-GRU	Random	N+S	$0.617_{0.015}$	$0.746_{0.042}$	$0.360_{0.054}$	$0.745_{0.024}$
Bi-GRU	MIMIC	N+S	$0.708_{0.014}$	$0.796_{0.027}$	$0.524_{0.023}$	$0.803_{0.016}$
ALARM	MIMIC	S	$0.326_{0.025}$	$0.607_{0.039}$	$0.065_{0.032}$	$0.307_{0.021}$
ALARM	MIMIC	N-S	$0.534_{0.041}$	$0.598_{0.027}$	$0.245_{0.088}$	$0.758_{0.038}$
ALARM	MIMIC	N+S	$\mathbf{0.766}_{0.028}$	$\mathbf{0.818}_{0.029}$	$\mathbf{0.661}_{0.061}$	$\mathbf{0.818}_{0.021}$

Table 3: Results for our ablation studies showing *macro-averaged* F1 as $\text{mean}_{\text{standard deviation}}$ of 5 runs with different random seeds (all models are trained with per-label attention). N data is the NHS GGC dataset and S is the synthetic dataset. "All" combines the classes "negative", "uncertain" and "positive". Bold indicates the best model for each metric.

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
2. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., Eisenstein, J.: Explainable prediction of medical codes from clinical text. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1101–1111. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1100
3. Wood, D., Guilhem, E., Montvila, A., Varsavsky, T., Kiik, M., Siddiqui, J., Kafiabadi, S., Gadapa, N., Busaidi, A.A., Townend, M., Patel, K., Barker, G., Ourselin, S., Lynch, J., Cole, J., Booth, T.: Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). In: Medical Imaging with Deep Learning (2020), https://openreview.net/forum?id=UFnWZTbM5t