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

Patrick Schrempf¹,², Hannah Watson¹, Shadia Mikhael¹, Maciej Pajak¹, Matúš Falis¹, Aneta Lisowska¹, Keith W. Muir³, David Harris-Birtill², and Alison Q. O’Neil¹,⁴

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                 | 0.0022±0.0001             |
| CAML [2]                     | 1,021,176   | 250.43            | 0.0090±0.0008             |
| Bi-GRU                       | 2,889,852   | 1113               | 0.0066±0.0003             |
| Bi-GRU + single attention    | 3,371,132   | 12055              | 0.0062±0.0003             |
| Bi-GRU + per-label attention | 3,401,852   | 3761               | 0.0109±0.0004             |
| BERT                         | 109,577,596 | 1115               | 0.0565±0.0025             |
| BioBERT                      | 109,577,596 | 927171            | 0.0575±0.0008             |
| ALARM + softmax              | 109,458,556 | 911243            | 0.0590±0.0013             |
| ALARM + per-label attention  | 125,233,276 | 1448375           | 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 mean±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    | 0.5840.022 | 0.4960.089 | 0.2040.031 | 0.6420.012 |
| Bi-GRU  | MIMIC     | N-S  | 0.9080.004 | 0.9560.004 | 0.4270.058 | 0.9270.004 |
| Bi-GRU  | Random    | N+S  | 0.8930.002 | 0.9620.008 | 0.4320.033 | 0.9030.002 |
| Bi-GRU  | MIMIC     | N+S  | 0.9210.003 | **0.9700.006** | 0.5730.011 | 0.9320.004 |
| ALARM   | MIMIC     | S    | 0.5690.028 | 0.7250.062 | 0.1280.041 | 0.5310.028 |
| ALARM   | MIMIC     | N-S  | 0.9060.011 | 0.9440.005 | 0.5320.087 | 0.9230.010 |
| ALARM   | MIMIC     | N+S  | **0.9280.008** | 0.9650.004 | **0.6890.039** | **0.9360.008** |

Table 2: Results for our ablation studies showing *micro-averaged* F1 as mean 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    | 0.4000.039 | 0.5040.050 | 0.1060.029 | 0.5900.065 |
| Bi-GRU  | MIMIC     | N-S  | 0.5510.024 | 0.6230.024 | 0.2680.089 | 0.7610.026 |
| Bi-GRU  | Random    | N+S  | 0.6170.015 | 0.7460.042 | 0.3600.054 | 0.7450.024 |
| Bi-GRU  | MIMIC     | N+S  | 0.7080.014 | 0.7960.027 | 0.5240.023 | 0.8030.016 |
| ALARM   | MIMIC     | S    | 0.3260.025 | 0.6070.039 | 0.0650.032 | 0.3070.021 |
| ALARM   | MIMIC     | N-S  | 0.5340.041 | 0.5980.027 | 0.2450.088 | 0.7580.038 |
| ALARM   | MIMIC     | N+S  | **0.7660.028** | **0.8180.029** | **0.6610.061** | **0.8180.021** |

Table 3: Results for our ablation studies showing *macro-averaged* F1 as mean 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., Kük, M., Siddiqui, J., Kafibadi, 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