Hospital-specific domain adaptation improves BERT-based models' ability to classify neuroradiology reports

Siddharth Agarwal[1], David Wood[1], Benjamin A K Murray[1], Yiran Wei[1], Ayisha Al Busaidi[2], Sina Kafiabadi[2], Emily Guilhem[2], Jeremy Lynch[2], Matthew Townend[1], Asif Mazumder[3], Gareth J. Barker[4], James H. Cole[5], Peter Sasieni[6], Sebastien Ourselin[1], Marc Modat[1], Thomas C Booth[1,2,*]

1. School of Biomedical Engineering & Imaging Sciences, King's College London; 2. Department of Neuroradiology, King's College Hospital NHS Foundation Trust; 3. Guy's and St Thomas' NHS Foundation Trust; 4. Institute of Psychiatry, Psychology, & Neuroscience, King's College London; 5. Centre for Medical Image Computing, University College London; 6. Clinical Trials Unit, King's College London, *Correspondence to: Thomas C Booth, thomas.booth@kcl.ac.uk, +44 (0) 203 299 4828

Introduction

- Classifying radiology reports is an important step in creating labels for downstream imaging tasks (i.e. prose into a single number)
- Expert-level domain knowledge required, infeasible at scale
- Bidirectional encoder representations from transformers (BERT) [1] models can generate 100,000 labels in 30 minutes once trained [2]
- Pretrained with masked language modelling (MLM)
    - On Wikipedia entries and novels it learnt English syntax, grammar
    - On Radiology corpora it learnt specific 'Radiology' English, improving its ability (after subsequent finetuning) to classify reports [3]

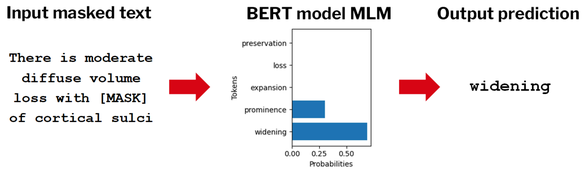


Figure 1: BERT models learn to predict masked words with masked language modelling (MLM), which often improves their performance in downstream tasks such as classification. The above example shows BERT MLM predictions after training on the KCH dataset.

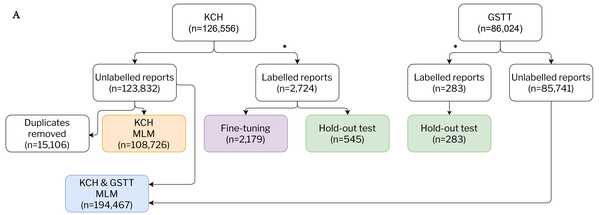
Hypothesis

Further gains can be made by repeating MLM at the deployment site i.e. hospital-specific domain adaptation

Aims

- Determine best overall training strategy and determine if there is incremental benefit from hospital-specific domain adaptation
- Compare BERT-based models to recent open-source LLMs

Methods



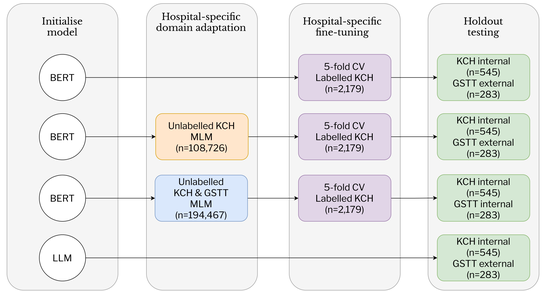
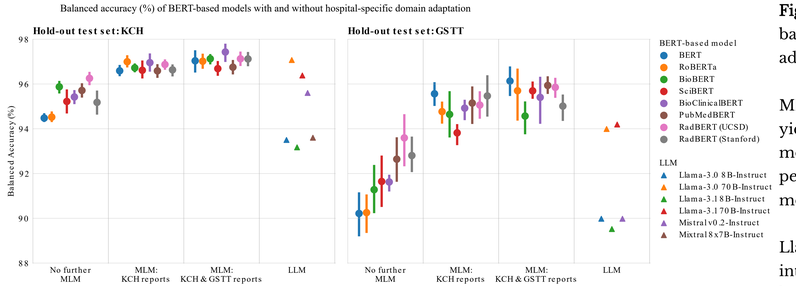


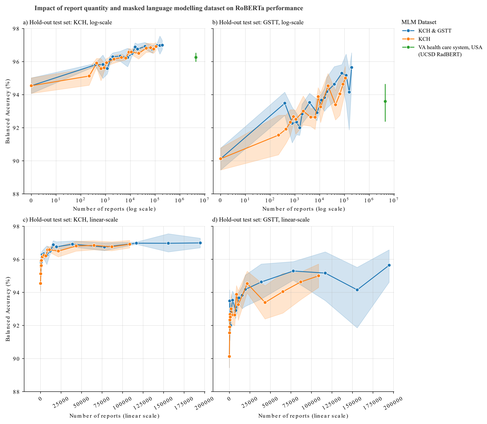
Figure 2A): Flow diagram of all adult MRI brain reports from KCH and GSTT, 2008 – 2018.

Figure 2B) Experimental setup. Each BERT model underwent hospital-specific domain adaptation on either no reports (Row 1), on unlabelled reports from KCH (Row 2) or on unlabelled reports from both KCH and GSTT (Row 3). Using transfer learning, each BERT model was then fine-tuned with 5-fold cross validation on labelled KCH reports to predict brain MRI reports as 'normal' or 'abnormal' (Column 3). Open-source LLMs were used as a benchmark, using a 6-shot chain of thought prompt (Row 4), and did not require training.

Ablation study: we examined the effect of choice and size of dataset by reducing the number of reports available for MLM and the number of reports for finetuning.

Results





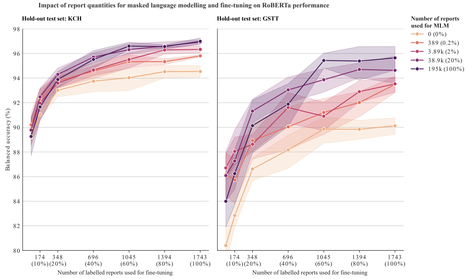


Figure 3 (left): Hold-out test set performance of BERT-based models with and without hospital-specific domain adaptation (mean and 95% CI are displayed).

MLM using all available unlabelled reports (194,467) yielded the highest balanced accuracies, each BERT-based model better than its base model. There was no performance accuracy difference between BERT-based models following MLM.

Llama-3.0 70B was the best performing LLM, with internal hold-out test set performance matching BERT-based models trained with MLM on 194,467 hospital-specific reports, but with slightly lower performance accuracy on the external test set

Figure 4 (left): Hold-out test set performance of RoBERTa [4], as a function of the number of reports available for MLM and the source of the reports (mean and 95% CI are displayed).

After controlling for base model (RoBERTa), language modelling method, and fine-tuning:
- Performance scaled log-linearly with the number of reports
- MLM with 194,467 local MRI brain reports yielded better performance compared to 4.42 million US general radiology reports; fewer than 15,000 local MRI brain reports were sufficient to achieve comparable performance.

Figure 5 (bottom left): Hold-out test set performance of RoBERTa [4], as a function of the number of reports available for fine-tuning and MLM (mean and 95% CI are displayed).

Discussion

- Hospital-specific domain adaptation consistently improved model performance and should be considered best practice
- Hospital-specific domain adaptation appears to make prior attempts at domain adaptation redundant (e.g. RadBERT) [3].
- Where training labels available we always recommend using BERT-based models with hospital-specific domain adaptation
- Where training labels scant, LLMs such as Llama 3.0 70B are a viable alternative

References

1) Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
2) Wood DA, Kafiabadi S, Al Busaidi A, et al (2022) Deep learning to automate the labelling of head MRI datasets for computer vision applications. Eur Radiol 32:725–736. https://doi.org/10.1007/s00330-021-08132-0
3) Yan A, McAuley J, Lu X, et al (2022) RadBERT: Adapting Transformer-based Language Models to Radiology. Radiology: Artificial Intelligence 4:e210258. https://doi.org/10.1148/ryai.210258
4) Liu Y, Ott M, Goyal N, et al (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach