Health system-scale language models are all-purpose prediction engines

Lavender Yao Jiang, et al.

New York University

September 15, 2023

Presented by Angel Huang

Introduction

Purpose

- Existing structured data-based clinical predictive models have limited use in everyday practice due to complexity in data processing, model development, and deployment – i.e. 'the last-mile problem'.
- Unstructured clinical notes can enable training of clinical language models and be used as all-purpose clinical predictive engines with low-resistance development and deployment.

Intuition: Using clinical notes to train an all-purpose clinical predictive engine **NYUTron** and fine-tune it to specific clinical and operational predictive tasks.

Potential Applications

- 30-day all-cause readmission prediction
- in-hospital mortality prediction
- comorbidity index prediction
- length of stay prediction
- insurance denial prediction
- My work: autism prediction

Background

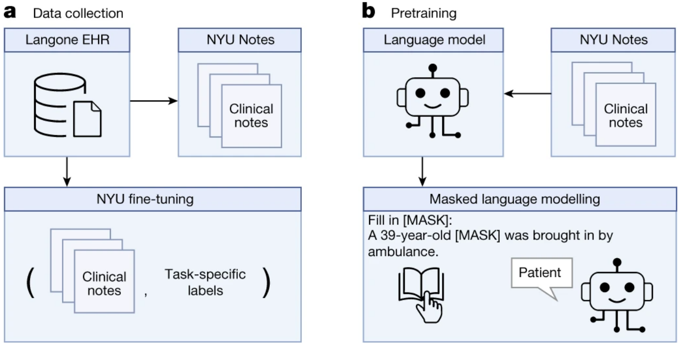
Clinical

- Medical decisions are based on information scattered across various records: medical history, laboratory, imaging reports, etc. Physicians integrate these into notes to summarize patient care.

BERT – Bidirectional Encoder Representations from Transformers

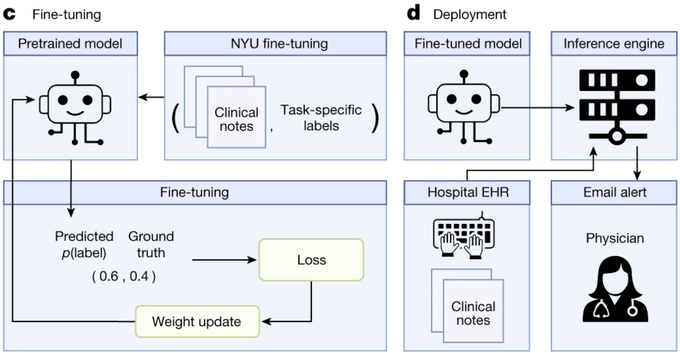
- BERT is a deep learning model based on Transformers, which processes any given word in relation to all other words in a sentence, the weightings between them are dynamically calculated based on their connections.
- BERT is able to process text "bidirectionally" and have access to both past and future tokens at learning time.
- BERT has enjoyed unparalleled success in NLP thanks to two unique training processes: Masked Language Modeling and Next Sentence Prediction.
- Masked Language Modeling: randomly masks words and trains the model to fill in the masked word correctly – words are defined by their surroundings, not by a pre-fixed identity (like work2vec).
- Next Sentence Prediction: predict whether two given sentences have a logical, sequential connection or their relationship is simply random – teaches BERT to understand longer-term dependencies across sentences.

Methods



- Two types of datasets:
  1. NYU notes: 10 years unlabeled clinical notes (radiographic reads, history and physicals)
  2. Five task-specific labeled clinical notes
- Pretrain NYUTron (BERT-like LLM) using MLM task until the validation loss plateaued

Methods



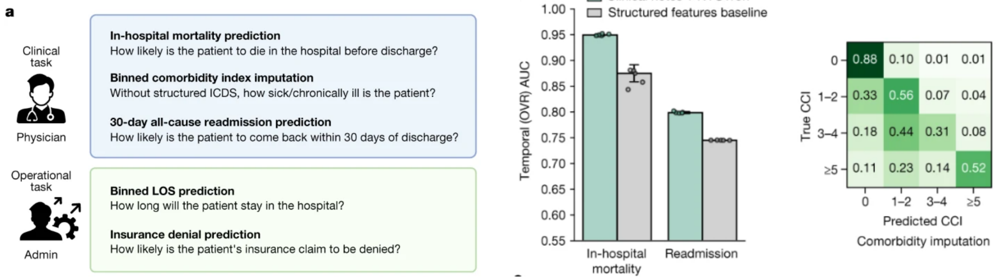
- Fine-tune on specific task, validate on retrospective data.
- Deployed best model to a high-performance inference engine that interfaces with EHR to read discharge notes.
- Deployment enabled real-time LLM-guided inference at the point of care.
- Validated performance on 30-day readmission prediction in a real-world environment

Implementation Details

- Language: SQL and Python, open-source packages with modification
- 387,144 patients, 7.25 million notes. Training: validation: test = 949:50:1 (patient-level leakage?)
- Temporal test set: Clinical notes sampled from the future of the training data (resembles the deployment scenario)
- Use one-vs-rest (OVR) AUC to evaluate the performance of multiclass classification
- The model is a relatively small LLM with <1 billion parameters: 12 hidden layers with dimension 768, with 12 attention heads per layer.
- Zero Redundancy AdamW optimizer (an improvement over the Adam optimizer) with a constant learning rate.
- Pretraining used 24 NVIDIA A100 GPUs with 40 GB of VRAM for 3 weeks, fine-tuning used 8 A100 GPUs for 6 hours per run.
- Deployment: Models were deployed utilizing a modified version of NVIDIA's TRITON Inference Server.

Experimental Results

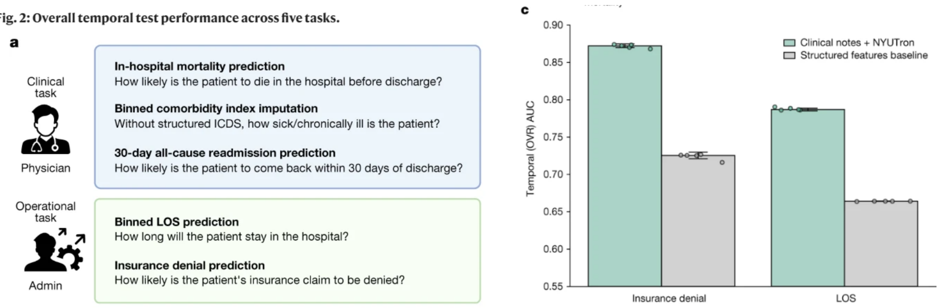
Fig. 2: Overall temporal test performance across five tasks.



- Performance: 7% and 5% improvement in mortality and readmission AUC.
- Structured baseline: XGBoost tree model
- Charlson comorbidity index (CCI) imputation task: 22% of data lacked chronic disease CCI score.
- Impute grades of severity (0, none; 1-2, mild; 3-4, moderate; $\geq$ 5, severe). No structured features were available.

Experimental Results

Fig. 2: Overall temporal test performance across five tasks.



- LOS was discretized into 4 quantiles.
- Performance: 12% and 15% improvement in insurance denial and LOS AUC.
- NYUTron can also predict different types of denials from both admission and discharge notes with similar performance.

Experimental Results: Detailed analysis on readmission

Additional evaluations

- Compare to human baseline (6 attending physicians predicting readmission of 20 random patients) – NYUTron performs better in TPR and F1 score
- Evaluate scaling properties with different numbers of fine-tuned data points – NYUTron scales better (same AUC at 100 to 1000 examples, but NYUTron improves with more data but XGB plateaued)
- NYUTron performs better than the randomly initialized LLM model and non-clinically pretrained models – 1/10 data needed to achieve 0.75 AUC.
- NYUTron generalizes better from fewer examples compared to non-clinically pretrained models – NYUTron performs better at 1000 examples, but the advantage disappeared as the number of fine-tuning examples increased.
- Model pretrained in one hospital can be generalized in another hospital (within NYU) – local fine-tuning yield best performance, but across-site only drops 0.01-0.02 in AUC.
- Prospective trial: deployed model to interface with EHR – AUC of 0.79.
- Clinical impact: Qualitative evaluation of 100 readmitted cases: they are clinically meaningful, preventable readmissions (50% of the unplanned readmission).

Conclusions

NYUTron is a health system-scale LLM for clinical use. With 1 pertaining dataset and different fine-tuning datasets, its performance was demonstrated on both clinical and operational tasks.

Insights

- Pretraining with in-domain clinical text is beneficial.
- High-quality datasets for fine-tuning are more valuable than pretraining.
- Any structured data algorithm can be conceptualized and rapidly prototyped within this framework.

Strength

- Evaluated the value of pre-training on clinical text, including sample efficiency, and generalizability across multiple sites.
- Tested in a deployment scenario and evaluated clinical impact.
- "Seamless integrated with existing medical workflows" in a live healthcare environment.

Conclusions

Concerns

- Substantial amount of computing time required, although "out-of-domain models can be highly performant when combined with in-domain fine-tuning"
- Structured baseline models didn't use many features, so the performance of baseline might be able to improve a lot.
- All 5 tasks are based on single note prediction. Not sure how to implement this to multiple notes prediction problem (eg. predict the probability of autism using all outpatient notes)
- Performance heavily depends on the quality of the physician's notes (including their opinions). Might perpetuate physician's bias.
- Fine-tuning still seems labor/computing-intensive, not sure if it solves the "last-mile problem".

Recommendations

- Seems like a scalable clinical/operational decision tool once developed.
- We can start with out-of-domain LLM with in-domain fine-tuning for small research projects.
- They have a well-structured Github repo that would be helpful.