Denis Jered McInerney

Research Statement

Research Experience

My research focuses on using machine learning to help clinicians discover important information in electronic health records (EHR). Because EHR interfaces are currently much better at collecting information than properly displaying it, the massive amount of unstructured information in EHR goes severely underutilized, even though it can be a crucial source of context for clinicians when treating patients. This gap motivates the idea that if only we could surface the right information from the EHR to clinicians at the right time, it could significantly reduce errors in patient care while potentially decreasing the already heavy workload on clinicians. Through three projects, I worked on developing methods to retrieve relevant sentences from a patient history, align text with local regions in images, and extract high-level features from notes, and currently, I am expanding on this work to prevent specific diagnostic errors with key evidence.

Query-focused Retrieval. Under the hypothesis that certain sentences buried in a patient's history could prove valuable to radiologists when looking at an image, this work set out to retrieve these sentences in response to a clinician's natural language query. Unfortunately, labels for extractive query-focused summaries do not exist for EHR. Annotating a training dataset for this task would be extraordinarily costly given the intensity of reading through the whole medical record of a patient for just one annotation and the need for the annotator to have clinical expertise. Instead, we proposed to use the eventual diagnosis of the patient, approximated by future ICD codes, as distant supervision with the intuition that sentences that are more informative of predicting a particular diagnosis should be relevant to that diagnosis. A manual evaluation with clinicians found that using the ICD code description text, not just the one-hot encodings, and incorporating the code hierarchy were important factors in achieving a model with high AUROC for sentence relevance scores.

Localized Multimodal Alignment. Next, we aimed to see if a similar logic could apply in identifying parts of an image that are relevant to queried text. Even if no localization labels are used for training, do the parts of the image most informative of some prediction, for example whether or not the text goes with the image, identify relevant regions? Existing work already created multimodal architectures with contrastive training objectives that train the model to make this very prediction (of whether an image and text correctly paired) as a pre-training objective. Further, it is popular to show heatmaps that imply that the regions picked out by the internal attention mechanisms of these models do align with intuition, but so far no one has critically evaluated this. Our work showed that while a model may produce reasonable heatmaps at first glance, changes in text do not intuitively change the heatmap. We further explored techniques designed to mitigate this undesirable behavior and showed that allowing the model to "not attend" to the image at all when the text does not align with the image is a promising direction.

Crafting High-level Variables with LLMs. Instruction-tuned LLMs have become extremely useful for zero-shot predictions, but it is difficult to deploy systems that use them in healthcare because they are black boxes. However, linear models are commonly used in healthcare because, like the models in the previous two projects, they allow inspection of how input features inform a particular classification decision. The downsides are two-fold: performance is lacking compared to more complex neural models and features are either too low-level to be meaningful or too high-level to be extracted manually. In this work, we proposed to use LLMs to extract complex high-level features that clinicians can specify with natural language and extract in real time. Then we use these features as input to simple linear models that offer the much needed interpretability. Our analysis shows that feature extraction and downstream classification (even with very few features) are reasonably successful, but most importantly, evaluation with a clinician reveals that the linear models' learned weights align with expert clinical judgement. The model can even retrieve features annotated as relevant to particular tasks in spite of not being trained to do so.

Current Work: Preventing Diagnostic Errors with LLM-retrieved Evidence. My current work focuses on surfacing and summarizing key evidence in EHR with LLMs that can prevent diagnostic errors when there is uncertainty. Specifically, we target three broad diagnoses that are common to the ICU, commonly misdiagnosed, and can result in serious consequences if misdiagnosed: Pneumonia, Pulmonary Edema, and Cancer. Using evidence retrieved with LLMs, we predict diagnoses while maintaining transparency in the way we aggregate the effect of each snippet of evidence. This allows us to 1) sort the retrieved evidence based on which evidence most impacts the aggregated prediction and 2) show which direction each individual piece of evidence "votes" on each condition. In an ongoing iterative process, we are alternating between refining the reliability and intuitiveness of our models and improving our annotation collection scheme to more closely resemble the envisioned use case—ICU clinicians using the system as an interface to alert them to important evidence when quickly assessing a patient. This project is the culmination of my previous work, collaboration with colleagues in the lab, and years of conversations with clinicians, and initial results demonstrate the potential of this technology.

Research Goals

I am passionate about working in areas with lots of unstructured data because this data tends to be vastly underutilized and there is a huge potential for immediate impact by simply getting the right pieces of information into the hands of domain experts. These application areas have the added benefit of illuminating the interesting and intellectually stimulating problems that need to be solved by fundamental research in machine learning and natural language processing.

For example, in healthcare, there is often too much information per example to handle: patient notes from potentially years of visits, fine-grained vitals, test results, and multiple radiology scans (potentially with multiple modalities and three dimensions). It can be difficult to combine and integrate this information in a computationally feasible way. When doing so, one often has to consider complex relationships between data within an instance, e.g. radiology scans are accompanied by reports, different note types might each tell a cohesive picture of one aspect of a patient, and everything has an important temporal relationship. Creating models that can handle longer inputs, hierarchical models, or models that simplify or structure the input can help to tackle these issues.

Models that use such high dimensional input are often prone to overfitting or can have difficulty with the distribution shifts (either through time or when transferring to a different source of data) that are common in real-world settings. This also requires technical innovation in safe, fair, and reliable generalization through better modeling or training techniques. We also need to develop more strategies that allow models to be easily trained in or adapted to the new distribution without an infeasible amount of annotation effort. In particular, I think it is a compelling use of LLMs and other pretrained models to extract structured targets from unstructured data in the "future" in time-series settings when structured information is unavailable or noisy.

Finally, the most immediate impact can be made by making models with which in-domain experts can actually collaborate. The simplest way to allow collaboration is to develop algorithms that point out underlying and succinct features of the data that experts can understand. However, to understand what is actually *useful* to show, we need to work with experts from the beginning. For example, clinicians we worked with found that it was more helpful to summarize information relevant to a prediction rather than make the prediction accurately. I am particularly interested in the idea of finding and summarizing temporal relationships in sequential data (e.g. the progression of an illness). In addition, clinicians may not be focused on making any particular prediction but instead on managing and making sequential decisions in the presence of uncertainty. Reinforcement learning can often be a useful tool in these settings, and I am interested in exploring how agents can be trained by using LLMs to model reward functions or simulate environments. These research goals offer ways to truly positively impact healthcare and other industries by augmenting domain experts rather than replacing them.